



FEB 08 2000

CENSUS 2000 DECISION MEMORANDUM NO. 97

MEMORANDUM FOR John H. Thompson
Associate Director for Decennial Census

Through: Preston J. Waite *PJW*
Assistant to the Associate Director for Decennial Census

From: Howard Hogan *Howard Hogan*
Chief, Decennial Statistical Studies Division

Subject: Overcounts for the Census 2000 Accuracy and Coverage
Evaluation Survey

Contact Person: Rick Griffin, Decennial Statistical Studies Division, Room 2500,
Bldg. 2, (301-457-4227)

I. Background

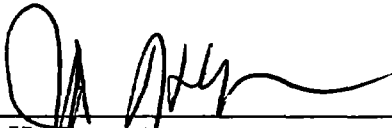
Census 2000 Decision Memorandum No. 90, Subject: Overcounts in Census 2000 Dress Rehearsal, presents a decision for dealing with poststrata with estimated overcounts. A decision was made for the Dress Rehearsal to stop selecting person records for subtraction when all imputed person records in a group had been used (See Decision Memorandum No. 90 for more details). This decision will be changed for Census 2000 as explained in this decision memorandum

II. Census 2000 Accuracy and Coverage Evaluation

For Census 2000 there will be no sampling for Nonresponse Follow-up or Undeliverable as Addressed Vacant units. Thus the proportion of the final count that is imputed persons will be much lower than for the Dress Rehearsal. To create an internal adjusted file for the 1990 Post Enumeration Survey, records for all persons (persons enumerated as well as persons imputed) in all poststrata were eligible for inclusion in the special coverage correction category. Not allowing records for enumerated persons in the coverage correction category would have resulted in an internal adjusted file with counts substantially different from the Dual System Estimates in poststrata with estimated overcounts.

If any overcounts are estimated for a particular poststratum for the Census 2000 Accuracy and Coverage Evaluation Survey (A.C.E.), the census counts for this particular group must be corrected to reflect the estimated overcount. The methodology for the A.C.E. survey will accomplish this by creating statistical records based on both enumerated and imputed data within the poststratum. These records will then be assigned a weight of -1 and included in the census data files in a special coverage correction category. This is in addition to the records that include the reports on enumerated and imputed individuals. When the census data are tabulated, the statistical records with the negative weights will be added to the census counts to incorporate the estimated overcount into the final results. Note that under this procedure no reported data for any individual will be removed from the Census 2000 data files.

I concur with the decision to allow all persons, enumerated and imputed, to be eligible for replication in the special coverage correction category with a weight of -1 for the Census 2000 A.C.E. Survey.



John H. Thompson
Associate Director for Decennial Census

FEB 08 2000

Date



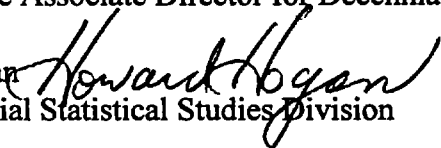
FEB 22 2000

CENSUS 2000 DECISION MEMORANDUM NO. 100

MEMORANDUM FOR John H. Thompson

Associate Director for Decennial Census

Through: Preston J. Waite 
Assistant to the Associate Director for Decennial Census

From: Howard Hogan 
Chief, Decennial Statistical Studies Division

Subject: Service-Based Enumeration in Census 2000: Multiplicity Estimation

Contact Person: Rick Griffin, DSSD, 2500/2, x4227

This memorandum describes changes in the statistical methodology used during the Service Based Enumeration (SBE) operation and documents the Census Bureau's decision to exclude data resulting from this methodology in the census counts generated for the apportionment of Member of Congress among the states. These data will be included in the adjusted counts produced after the completion of the Accuracy and Coverage Evaluation survey, which the Census Bureau will make available in a form that allows states to use them for redistricting purposes. The more accurate counts can also be used for determining the allocation of federal funds, and for ongoing statistical and programmatic purposes.

I. Introduction

The Service-Based Enumeration (SBE) operation is the Census Bureau's primary program for enumerating people with no usual residence. The Census Bureau designed this operation to enumerate people at service locations that primarily serve people without usual residence, such as emergency and transitional shelters, soup kitchens and regularly scheduled mobile food vans.

As part of the compromise on sampling reached by the Administration and Congressional leaders, the Census Bureau agreed that statistical sampling would not be used at the Columbia, South Carolina site during the Census 2000 Dress Rehearsal. At the Sacramento, California site, the Census Bureau tested statistical sampling as part of the Census 2000 plan in place at the time. Consequently, two different methodologies were used to include people without a usual residence in the Census 2000 Dress Rehearsal.

In Columbia, SC the Census Bureau visited emergency and transitional shelters on April 20, 1998. A two member enumeration team enumerated people at most shelters using Individual Census Reports. Larger sites had more than two enumerators. At soup

kitchens enumerators conducted personal interviews using an Individual Census Questionnaire. After an unduplication process was complete, the count was assumed to be the total number of people enumerated at these facilities.

In Sacramento, CA all field procedures and questionnaires were identical to the ones used in Columbia, SC, but a multiplicity estimator based on responses to the usage questions was applied. A usage question asks respondents how many times they used service facilities in the past week. This estimator accounts for people who use services, but were not present on the day of the enumeration.

The Census Bureau will not use the multiplicity estimator to determine the apportionment counts. In order to obtain a more accurate count of persons without usual residence, the Census Bureau has decided to use the multiplicity estimator to produce Census results for all other uses except apportionment.

II. Change in SBE Multiplicity Estimation for Census 2000

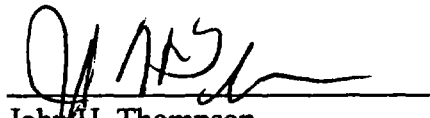
For the Census 2000 Dress Rehearsal multiplicity estimation any SBE person who did not respond to the usage question was given a weight of zero. Enumerated persons who did respond to the necessary usage question had their multiplicity weight multiplied by a noninterview factor to account for those persons with a weight of zero. As a result we discarded demographic data actually collected for persons who did not respond to the usage questions. In addition this zero weighting caused problems with presentation of results in terms of persons added due to SBE multiplicity estimation.

For Census 2000, we will impute responses to the usage questions prior to multiplicity estimation. Thus, after unduplication and controlled rounding all SBE persons will be included on the file used for all purposes except apportionment at least once. The number of persons added due to multiplicity estimation will be the adjusted SBE count minus the unadjusted SBE count.

III. Decision

The population without usual residence is very transient (by definition). Thus, using the traditional methodology would require numerous visits to the service locations to obtain a reasonable count. Census 2000 will use the same data collection procedures (one visit) that were used in the Dress Rehearsal. The multiplicity estimator accounts for those persons without usual residence who use services but who were not present on the day of enumeration. In addition since Census 2000 Dual System Estimation (DSE) excludes all Group Quarters persons including SBE persons, SBE multiplicity estimation does not interfere with the critical path for DSE. Multiplicity estimation should be used for all purposes except for apportionment for Census 2000.

I concur with the use of the multiplicity estimator for Census 2000 for all purposes except apportionment.

A handwritten signature in black ink, appearing to read 'JH Thompson', written over a horizontal line.

John H. Thompson
Associate Director for Decennial Census

FEB 22 2000

Date



UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

MASTER FILE

August 19, 1999

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-2

MEMORANDUM FOR Howard Hogan
 Chief, Decennial Statistical Studies Division

From: Donna Kostanich *DK*
 Assistant Division Chief, Sampling and Estimation
 Decennial Statistical Studies Division

Prepared by: Inez Chen *IC*
 Estimation Staff
 Decennial Statistical Studies Division

Subject: Census 2000 Overview of Unclassified Estimation

Introduction

The purpose of this memorandum is to provide a high-level description of how we plan to impute data missing from census housing unit records. This procedure is referred to as unclassified estimation, since we are concerned with only missing housing unit status and missing population count for occupied housing units. Other missing data are handled during the Content, Edit, and Imputation procedure which occur after unclassified estimation is complete. Only the unclassified estimation is required for the apportionment counts due December 31, 2000.

After many stages of census operations, such as mail list development, update/leave, and list/enumerate operations, postal verification check, new construction program, and other late adds operations, a final list of housing units existing on census day is established. At the end of follow-up activities and data capture processing, some census housing unit records will not contain information on the number of persons or will not contain information on whether the census housing unit is occupied, vacant, or delete. The source of these omissions may be from respondents not providing correct or timely information or from any unanticipated operational obstacles.

The unclassified estimation is designed to fill in missing housing unit status and the number of persons for any occupied census housing unit without household size. The operation will be done concurrently as part of the creation of Census Unedited File (CUF), on a flow basis by Local Census Office (LCO). Unclassified estimation is the last process to complete the CUF.

Other CUF creation activities such as merging the Decennial Master Address File with the Decennial Response File for establishing census housing unit records will have been completed prior to the unclassified estimation operation.

Methodology

Under the assumption that housing unit status and number of persons living in a housing unit are more similar in a nearby neighborhood than a far away community, the nearest-neighbor hot deck method will be used. This means that the data from the closest available neighbor will be used to fill in the data for the unclassified housing units. Geographical closeness of housing units is determined by sorting all housing units and group quarters within a tract by block number, street name, and house number. Based on the sorted sequence, backward and forward searches will be conducted to find a donor for an unclassified unit¹. The nearest available classified unit in either direction will be used as a donor to fill in the data for the unclassified unit.

To keep possible bias from this operation to a minimum, potential donors (classified units) and donees (unclassified units) will be grouped into the following four groups:

- Early response (before nonresponse followup) and not in a multi-unit structure
- Early response and in a multi-unit structure
- Late response and not in a multi-unit structure
- Late response and in a multi-unit structure

The nearest-neighbor hot deck will be done separately for each of these four groups. One restriction is that the donor for an occupied unclassified unit will be an occupied classified unit. If possible, the donor will be in the same tract as the donee.

Summary

For the Dress Rehearsal, the unclassified estimation was conducted in conjunction with Nonresponse Follow-up (NRFU) and Undeliverable-as-Addresses Vacant (UAA) estimation. For 2000, sampling for NRFU and UAA is not allowed, thus the estimation is also eliminated from the program. For Census 2000, production for unclassified estimation will start in mid-September 2000 and be complete by October 3, 2000. A final working draft specifications

¹ A second nearest donor will also be identified for evaluation purposes.

for the operation will be available around mid-September of 1999. Estimation staff of the DSSD will work closely with programmers to develop and test software for the operation.

Please direct any questions or concerns to Inez Chen on (x4106).

cc: DSSD Census 2000 Procedures and ^{R.D.} Operations Memorandum Series Distribution List
Statistical Design Team Leaders
C. Hoang (DSSD)



DEC 29 1999

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-17

MEMORANDUM FOR

Howard Hogan
Division Chief
Decennial Statistical Studies Division

From:

Donna Kostanich *DK*
Assistant Division Chief for Sampling and Estimation
Decennial Statistical Studies Division

Prepared by:

Roger Shores *R.S.*
Long Form and Variance Estimation Staff
Decennial Statistical Studies Division

and

Carl Durant *CD*
Estimation Staff
Decennial Statistical Studies Division

Subject:

Treatment of Late Census Data for Accuracy and Coverage Evaluation
Estimation

I. INTRODUCTION

The Accuracy and Coverage Evaluation (A.C.E.) housing unit and person matching operation requires the identification of housing units and persons in the E-sample. The E-sample consists of all census persons residing in housing units that were selected in the sampling process. The Census Unedited File (CUF) is created and used to identify the housing units and persons in the E-sample. Essentially, the E-sample is a subset of the housing units and persons on the CUF. Therefore, persons in housing units added to or deleted from the census after the CUF is created will not be included in the E-sample. The CUF is to be created and delivered by the start of A.C.E. operations on September 21, 2000.

Late changes to the Census inventory, whether they are additions to or deletions from the CUF after it is delivered, could have significant implications for A.C.E. estimation. It is important to understand, however, that census plans call for all housing units and persons to be identified in time for CUF delivery. It is fully expected that there will be no late census data. The procedure

recommended for handling such data constitutes, therefore, a contingency plan that is quite unlikely to require implementation. **The recommendation itself is that late census data not be included in any census counts for the calculation of the DSE. Coverage factors, on the other hand, are calculated with late census data included in the final census count.** This plan should be straightforward to implement, while producing a DSE whose expected value would be similar to the expected value of the DSE with the late census data included.

For the purposes of this document persons who were included on the CUF and later removed are defined as late deletes. Persons added to the census after the CUF is delivered are defined as late adds. If late adds and late deletes do occur, they will cause the Dual System Estimate(DSE) to be inaccurate by making it too high in the former case and too low in the latter. The DSE formula is

$$DSE = C \cdot \frac{CE}{E} \cdot \frac{P}{M}, \quad (1)$$

where

C = Final census count (excluding non-data defined persons)

CE = estimate of census correct enumerations

E = E-sample estimate of total population

P = P-sample estimate of total population

M = estimate of P-sample matches to census enumerations

The rest of this memorandum describes the alternatives that are available for dealing with late census data.

II. LATE CENSUS DATA

DELETES

Housing units and persons erroneously included in the E-sample result in a lower correct enumeration rate and thus an underestimated DSE. The solution is to simply remove them from the E-sample. Doing so will give the value of the DSE that would have been calculated had those units and persons not been wrongly included in the E-sample. The removal should have little, if any, effect on other components of the DSE formula. The match rate, in particular, should not change significantly. The P-sample will presumably contain very few people who were erroneously included and matched to an erroneously included person in the E-sample. All that need be done, therefore, is to obtain a listing of housing units and persons deleted from the census after the CUF is created and remove them from the E-sample. The DSE formula for this alternative is

$$DSE = C \cdot \frac{CE}{E - E_D} \cdot \frac{P}{M}, \quad (2)$$

where E_D = late census deletes.

It should be noted that this procedure assumes that all census deletes are true erroneous enumerations. This is significant because incorrectly deleting persons from the E-sample would cause the match rate to be too low, since presumably those people would match to P-sample people. The result would be an upward biasing effect on the DSE.

ADDS

The situation for the late adds is not as straightforward as that for the deletes. Housing units and persons added or data captured after September 21, 2000 will not be in the E-sample because the CUF will have been created by that date. The people in those units will, however, be in the P-sample. The discrepancy causes the match rate to be lower than it would have been had those units and persons been included in the E-sample. The result is that the DSE is overestimated.

There are three options under consideration for dealing with late adds:

1. Do nothing.

This means using formula (1) to calculate the production DSE.

2. Include the late adds in the E-sample and match them to people in the P-sample.

The late adds would be considered the same in terms of the matching process as previously included members of the E-sample. The resulting DSE would be the one that would have been calculated if the late adds had been in the E-sample at the time of the initial matching. The DSE formula is then

$$DSE = C \cdot \frac{CE + CE_A}{E + E_A} \cdot \frac{P}{M + M_A}, \quad (3)$$

where CE_A = correctly enumerated late adds,
 M_A = late additions that match to the P-sample, and
 E_A = late census adds.

3. Put the late adds in the E-sample as erroneous enumerations.

Doing so will lower the correct enumeration rate and thereby offset the low match rate. The DSE formula for this option is

$$DSE = C \cdot \frac{CE}{E + E_A} \cdot \frac{P}{M}, \quad (4)$$

where E_A = late adds included in the E-sample. They are treated as erroneous enumerations.

Implementation of alternatives (2) and (3) requires possession of a list of the late census adds. The second alternative would clearly be the best of the three with respect to the DSE estimate and its variance, but would be quite difficult to implement because of time and resource constraints. The third alternative would have a similar effect on the DSE while being easier to accomplish, though at the cost of being slightly biased in some situations.

In addition to these three, there is a fourth alternative, one that applies to both late adds and late deletes. It is as follows:

4. Do not include late adds or late deletes in the E-sample data used for estimation.

In other words, do nothing with respect to late census data. **Do not** include late census data in the census count (the first component of the DSE estimate) used for DSE estimation. For adds, a lower census count will offset the effect of the low match rate on the DSE. For deletes, a larger census count will compensate for a lower correct enumeration rate. To compute coverage factors, proceed as usual by taking the ratio of the DSE to the census count. The census count used for this operation **must include** the late census data. As in the past, the census count will be a tally of the unadjusted detailed file. To implement this option, counts of census adds and deletes by demographic/tenure variables will be required for review purposes. The counts have been requested by DSSD in a July 16, 1999 draft memorandum¹ from Howard Hogan. The DSE formula for this alternative is

$$DSE^* = C_1 \cdot \frac{CE}{E} \cdot \frac{P}{M} \quad (5)$$

In practice, $C_1 = C + C_D - C_A$, with

C_D = late census deletes, and

C_A = late census adds.

The late deletes are added back to the census count, and the late adds are subtracted from it, to obtain the census count that existed at the time the CUF was delivered. This is the count that does not include any late census data.

The coverage factor can be written as

$$CF = \frac{DSE^*}{C} \quad (6)$$

where $C = C_1 - C_D + C_A$. The census count used to calculate the coverage factor is the final census count; therefore, starting with C_1 , the late deletes need to be removed, and the late adds put back in, to obtain it.

¹ Subject: "Review of Invalid Return Detection Business Case Analysis".

III. CONCLUSION

Handling the deletes should not be especially difficult. For the adds, the third alternative would seem a reasonable compromise between not doing anything to adjust the overestimate and a solution that is not practicable. However, the fourth alternative is more comprehensive, and does not require any modifications to the E-sample file and does not give the appearance of data manipulation. It does not call for anything beyond that needed to implement the third option. The only requirement is the production of two sets of census counts for A.C.E. estimation, one with and one without late census data. The Program Steering Committee for Statistical Design has recommended that this alternative be implemented if there is late census data.

cc:

DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List



MASTER FILE

January 12, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-18

MEMORANDUM FOR Howard Hogan
 Chief, Decennial Statistical Studies Division

From: Donna Kostanich *DK*
 Assistant Division Chief, Sampling and Estimation
 Decennial Statistical Studies Division

Prepared by: Alfredo Navarro *AN*
 Chief, Long Form and Variance Estimation Staff

Subject: Accuracy and Coverage Evaluation Survey:
 Targeted Extended Search Plans

1. Background

The 1990 Decennial Census used a Post-Enumeration Survey (PES) methodology to measure census undercount or overcount. The Census Bureau used the dual-system model to produce an estimate of total population. This model relies on classifying each person from the "true" population as being either included in the Census or not, as well as being included in the PES or not. The 1990 PES was conducted on a sample of blocks and used the dual-system estimator (DSE) to produce estimates of total population for various demographic groups. The DSE estimation rules were very simple. For example, a person was considered a census enumeration if he or she was tallied in the population count. This person or census enumeration was considered a correct enumeration if he or she was supposed to be included in the census. This is referred to as "Appropriateness" (Hogan, 2000.) On the contrary, a person counted in the census when he or she was not supposed to be included was termed an erroneous enumeration. A person was classified as a census omission if he or she was supposed to be in the census enumeration but was not. There was also a requirement for people to be counted in the "right location."

In the context of the 1990 Census PES, right location meant anywhere in the block where the reported housing unit address was located. In addition, the 1990 PES design employed the concept of search area. For the most part, the search area was defined as one ring of adjacent blocks. In rural areas, the search area was expanded to two rings of surrounding blocks. In List/Enumerate areas, the entire address register area (ARA) was searched. Using search areas, as long as a census (E-sample) person was counted in the correct block or in any of the blocks in

the ring(s) of surrounding blocks he or she was labeled as a correct enumeration. Persons in the P-sample found in the search area were treated as matches, that is, as not missed by the census. If these two effects are balanced the measure of net undercount is not affected. Failure to balance the two effects results in "balancing error".

The process of searching for census omissions and erroneous enumerations within the search area is very costly, labor intensive, and requires highly trained and skillful clerks (Hogan, 1993.) The process and resources required to repeat the 1990 PES extended search operation in the 2000 A.C.E. are extremely difficult to effectively operationalize. Limiting the amount of searching to block clusters with a potential high payoff will result in a more accurate and efficient search operation. The 2000 A.C.E. search operation differs from the 1990 PES search process in three ways. First, for the 2000 Census A.C.E. the concept of right location is more limited than the concept used in the 1990 Census PES. Secondly, the search itself will be targeted or limited to persons in whole household nonmatches, namely geocoding errors of exclusion and inclusion. Geocoding errors of exclusion affect the P-sample match rate (or the census coverage rate.) Geocoding errors of inclusion affect the E-sample erroneous enumeration rate. A third major difference is that the 2000 A.C.E. search operation will be sample based, that is, the extended search will be implemented for a subsample of the A.C.E. sample. The 1990 PES performed a search for all block clusters in sample. Targeted Extended Search or TES is the manner by which we will identify and select cases to be included in the extended search operation.

The implementation of the DSE does not require an extended search operation. The expectation of the DSE is not affected by the introduction of the search area concept. In other words, if the search area is limited to the PES block cluster¹ (as in the 1995 Census Test) the expectation of the DSE is the same as that under the 1990 PES search area definition. The motivation for using an extended search area definition is variance reduction. Allowing more cases to be matched and more census enumerations found in surrounding blocks result in a higher match and correct enumeration rate, respectively. This will yield a DSE with more precision or less variance.

This document describes the methodology for targeting, sampling, and operational issues associated with the A.C.E extended search plans for Census 2000. The concepts of "targeting" and "balancing error" are further discussed in Section 2. In addition, Section 2 describes search plans for 2000. It also gives some results from empirical research performed to compare alternative TES sample designs considered for implementation in Census 2000. Section 3 describes the targeting methodology criteria and sampling operations used for identifying and selecting the TES block clusters. Section 3 also describes the extended search operations for persons. The final section describes the TES estimator and highlights the effects of the TES on dual system estimation.

¹ A block cluster is a group of blocks and was the 1990 PES sampling unit. The average size of a block cluster was about 30 housing units.

2. Search Plans for 2000

The 2000 Census A.C.E. search operation differs from the 1990 PES in three areas, these are:

- a) search area definition
- b) amount of searching
- c) eligible people for searching

The search area for the 2000 A.C.E. will be limited to either just the sample block cluster or at most one adjacent block. An adjacent block is one that touches the cluster of sample blocks at one or more points. This definition includes the blocks that touch the corner of the block cluster. Results from empirical research using Census 1998 Dress Rehearsal data show that the additional benefits of using two rings of surrounding blocks are almost negligible (Wolfgang, 1999.) Additionally, the plan is to implement a sample based search operation. The targeting will be implemented in two phases. Specifically, the plan calls for targeting the extended search operation to 20 percent of the A.C.E. sample block clusters. The second phase limits the searching of census omissions and erroneous enumerations due to duplication within the adjacent block where the housing unit is found. The search will be concentrated in block clusters thought to have the biggest payoff in terms of variance reduction. Census geocoding errors affect both the census omission rate (or the P-sample match rate) and the census erroneous enumeration rate. Low correct enumeration and match rates have a negative impact on the reliability of the DSE. Block clusters with high concentration of census geocoding errors can be identified from the results of the initial housing unit matching and subsequent field follow up. These are A.C.E. block clusters with a large number of P-sample housing units (or addresses in the Independent Listing) not found in the E-sample (or the Decennial Master Address File) and referred to as whole household non-matches. These types of non-match are possibly census geocoding errors of exclusion. On the Census side, there are A.C.E. block clusters with a large number of census geocoding errors. These are referred to as census geocoding errors of inclusion. Results from the 1990 PES show that geocoding error is highly clustered. Slightly over 77 percent of the whole household nonmatches were concentrated in less than one-fourth of the PES sample block clusters. On the other hand, about 72 percent of the census geocoding errors were found in less than 3 percent of the PES sample block clusters (See Attachment - TES Empirical Research - Summary of Results, Tables 2 and 3.) It seems that this is a clear example of a Deming principle, the so called "80-20" rule which states that in most cases "80 percent of the benefits are realized by solving 20 percent of the problems."

2.1 Targeting Methodology

The proposed plan for the 2000 A.C.E. is to target the extended search in the surrounding blocks of 20% of A.C.E. block clusters. Based on the 1990 PES experience we developed a well defined targeting criterion that when applied will result in the selection of A.C.E. block clusters with superior payoff. From the 1990 PES we learned that one reason for census omissions and

erroneous enumerations was census geocoding error. This type of census omissions and erroneous enumerations will benefit the most from an extended search in surrounding blocks. *The criterion is to identify TES block clusters on the basis of independent listing unmatched housing units with a nonmatched census address. On the E-sample side the criterion is the number of housing units geocoded erroneously in the E-sample block.*

2.2 2000 A.C.E. TES Sampling Plan

An empirical simulation was designed and performed to assess the effect of alternative TES plans on the DSE and its variance (See Attachment.) We used the 1990 PES data base for the simulations. The results are conditional on the 1990 PES experience. Although it is important to note that there are many differences between the 1990 PES and the 2000 Census A.C.E., the simulation results provide the basis for discriminating between the alternative TES plans. The TES sampling plans simulated fall into two categories, these are:

- Certainty selection
- Combination of certainty and probability sampling.

Certainty samples ranging in size from 5 to 20 percent were simulated. All these samples yielded more reliable DSE's compared to not doing any search but with varying degrees of conditional bias. The reliability of the DSE based on a 20 percent TES certainty sample is very close to the precision of the 1990 DSE based on a full PES sample extended search. However, it is a very difficult task, perhaps impossible, to design a balanced certainty sample. To compensate for this, we developed several plans based on a combination of certainty and probability sampling. For these sampling plans, half of the TES sample was selected with certainty and the remainder was selected using a systematic sampling scheme. These sampling plans produced conditionally unbiased results and had almost no effect on the variance of the DSE. Based on the simulation results, we developed the following TES sampling plan for implementation in 2000.

Certainty Sample

- Five percent of clusters with the most census geocoding errors and independent listing address nonmatches.
- Five percent of clusters with the most weighted census geocoding errors and A.C.E. address nonmatches
- All relisted clusters in the P sample.

Probability sample

- A systematic sample from the remaining clusters with at least one census geocoding error or an independent listing unmatched address. The number of clusters in the sample will be determined so the total numbers of block clusters in sample is equal to 20 percent of the A.C.E. sample.

A.C.E sample clusters in List/Enumerate areas are out of scope for TES sample selection. List/Enumerate clusters will be handled through special procedures. Special procedures are being developed as needed for clusters with high person nonmatch and census geocoding error rates.

2.3 TES Sampling Operations

Results from the initial housing unit matching operation will be used to identify the TES sample. Housing unit matching consist of several operations. These are:

- Computer match - Addresses in the January 2000 Decennial Master Address File (DMAF) extract are computer matched to addresses in the A.C.E. Independent Listing.
- Clerical match - For this operation the search area is limited to the sample block cluster.
- Housing Unit Follow up - Results from the computer and clerical match operations are used to identify cases to go to the field for follow up. The goal of this operation is to create an accurate inventory of all housing units in the block cluster.
- After Housing Unit Follow up Coding - Using the information collected during field follow up, housing units are assigned one of several codes.

For TES sample selection we are interested in three types of housing units, these are:

- CI - The A.C.E. housing unit existed as a housing unit and is correctly geocoded in the block cluster. An address corresponding to the housing unit is not found in the census.
- UI - Not enough information to determine the match status of the housing unit with certainty.
- GE - The census housing unit existed as a housing unit but is incorrectly geocoded in the block cluster. The housing unit is a geocoding error.

See attached Figures 1 and 2 for a summary of housing unit follow up codes.

TES block clusters will be identified based on the number of housing units coded as GE, CI, and UI. The total number of housing units in these three categories will be obtained for each block cluster to identify the certainty clusters as described above. The sampling plan will be

implemented to select the TES probability sample. The search area is expanded only for these TES block clusters.

3. TES Operations

3.1 TES Housing Unit Operations

The search area is expanded for the block clusters selected for TES. A field visit is conducted in TES clusters to identify the housing units that exist in the surrounding blocks. This visit will be conducted during the A.C.E. person interview phase. During the TES field follow up visit the exact location of the housing unit is determined and recorded for use in the person matching phase.

Housing units added and deleted in TES sample clusters after the initial housing unit matching are a limitation for the implementation of the TES operations. The added housing units will not undergo housing unit matching, therefore we would not know which units are geocoding errors at the time of sampling. Geocoding errors will be identified during person follow up for this set of housing units. To identify which units actually exist in the surrounding blocks from the ones that exist outside the expanded search area we would need to add another field operation. The A.C.E. operation managers decided that this additional operation could not be implemented without putting at risk meeting key deadlines in the A.C.E. schedule.

This decision triggered a decision not to include in the TES the P-sample people in P-sample housing units whose address matched initially to a census housing unit but was later deleted from the final census inventory. As a result of the unit being deleted, the P-sample household becomes a whole-household nonmatched. This decision was made to avoid introducing balancing error.

3.2 TES Person Operations

The TES person definition is as follows:

- **P sample** - Persons in whole household nonmatches with no census address match. Partial household nonmatches or persons in whole household nonmatches with a matched census address are not included in the TES definition. Persons in whole household nonmatches with an initial census address match which is deleted by a later census operation are excluded from the TES definition.
- **E sample** - Persons in housing units that are recorded as census geocoding error during the initial housing unit matching and confirmed to exist in the surrounding blocks of the search area. These housing units are referred to as census geocoding error in the surrounding blocks and the people are coded GS.

3.3 P-sample Person Operation

During the extended search operation address matching is conducted in the search area, which is one ring of surrounding blocks. If the basic street address is found in a block in the search area, then person matching follows only in the block where the basic street address is located. If the persons are not found in the block then the P-sample people are nonmatches. If these persons are found and matched in the surrounding blocks they are treated as matches during dual-system estimation. Person searching is also performed for possible address matches.

3.4 E-sample Person Operation

During the housing unit extended search operations TES eligible housing units are searched in the surrounding blocks of the sample cluster. Housing units coded GE after the initial housing unit follow up are TES eligible housing units. Persons in housing units not found in the search area are coded as erroneous enumeration. Persons in housing units found in the surrounding blocks will undergo a duplicate search. A targeted duplicate search for persons will be conducted in the block where the housing unit should have been counted. If the housing unit is not duplicated, a search for persons is not performed. We are only interested in identifying person duplication due to census geocoding error. If the census record is found and the persons confirmed to be duplicated in the surrounding block, the E-sample persons are treated as erroneous enumerations for dual-system estimation.

4. TES Estimation

A simple expression for the DSE estimator is:

$$D\hat{SE} = (C-II) \left(\frac{CE}{N_e} \right) \left(\frac{N_p}{M} \right); \text{ where}$$

C- II = census data defined persons

CE = weighted estimate of correct enumerations from the E-sample

N_e = weighted estimate of E-sample persons

M = weighted estimate of P-sample matches

N_p = weighted estimate of P-sample persons

The definition used for a "TES Person" is someone who is only eligible to be matched or converted to a correct enumeration if TES is used in the person's block cluster, a trait that can be determined from the characteristics of the person's housing unit.

A general expression for estimating each of the DSE components is as follows:

$$\sum_{i=1}^n \sum_{j=1}^{n_p} w_{ij}^* m_{ij} x_{ij} + \sum_{i=1}^n \sum_{j=1}^{n_p} w_{ij}^* m_{ij} y_{ij} + \sum_{i=1}^n \sum_{j=1}^{n_p} w_{ij}^* t_{ij} m_{ij} z_{ij}$$

where;

i= cluster index

j= person index

n= number of block clusters in the A.C.E. sample

x_{ij} = 1 if the person is not a TES person, 0 otherwise

y_{ij} = 1 if the person is a TES person and is in the TES sample with certainty, 0 otherwise

z_{ij} = 1 if the person is a TES person and is in the TES systematic sample, 0 otherwise

m_{ij} = characteristic of interest, match, correct enumeration, E-sample person, or P-sample person

w_{ij}^* = weight used for estimation (includes inverse of the probability of selection for A.C.E., adjustment for household noninterview and missing data imputation)

t_{ij} = TES sampling weight, the TES systematic sample take-every

For example, the above formula is used to obtain the estimate of correct enumerations for each post-strata. We will use the same expression to get estimates for the number of matches, E sample population, and P sample population. Since the TES will be conducted for a sample of the A.C.E. sample we need to incorporate the TES sampling weight into the dual system estimation operations. All four A.C.E. components (CE, N_e , M, N_p) of the DSE reflect the TES probability of selection. This estimator is referred to as *double expansion estimator* (Kott and Stukel, 1997.) Thus, TES persons in blocks not selected in the TES sample are not included in the estimation of E and P sample population totals. These not sampled TES persons are accounted for by adjusting the weights of TES persons by the TES sampling weight. The estimator is efficient since it takes advantage of the correlation between the estimate of correct enumerations (P-sample matches) and the E-sample (P- sample) population total.

References

Hogan, Howard (1993), "The 1990 Post-Enumeration Survey: Operations and Results", JASA, 88, 1047-1058.

Wolfgang, Glen (1999), "Request for Dress Rehearsal Surrounding Block Files for A.C.E. Research", March 29, 1999, unpublished Census Bureau memorandum.

Kott, P.S. and Stukel, D.M. (1997), "Can the Jackknife Be Used with a Two-Phase Sample?", Survey Methodology, 23, 81-89.

Attachment

Targeted Extended Search Empirical Research - Summary of Results

The accompanying Table 1 illustrates the effect that various block cluster selection criteria had on the DSE and its variance. The table illustrates only the effect on the poststratified DSE of the total population, but a similar trend was observed for most population groups studied. Using the 1990 PES data, we simulated the DSE and estimated variance under a number of different selection strategies. The strategies are described as follows:

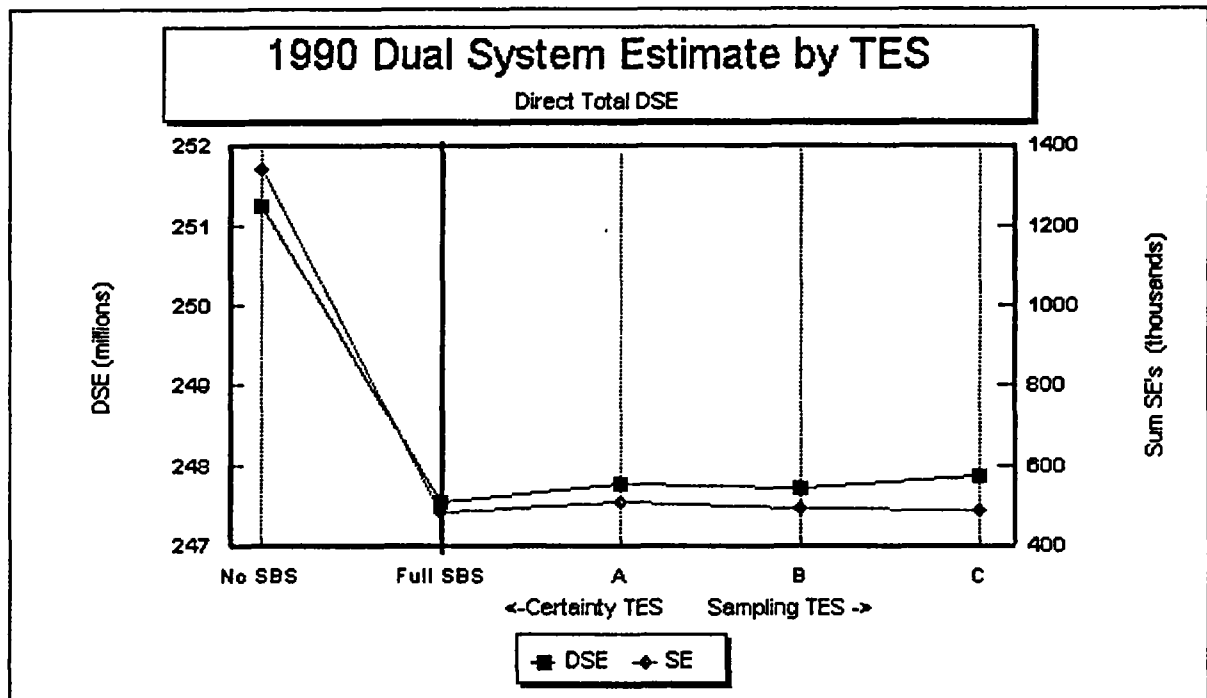
- Full Surrounding Block Search (SBS) or 100 percent TES sampling – the actual 1990 results. Because of slight differences between the files we had available and those used for 1990 publication not all data values are exactly the same as 1990 published data, but are in all cases close.
- No SBS or Zero percent TES sampling - we simulated undoing the effect of the 1990 SBS.
- Certainty plus sampling – Ten percent of block clusters were selected based on certainty selection only and then an additional 10% were selected by a sample of those not included with certainty and eligible for TES sampling. Various methods of selecting the certainty part are referred to as Plans A, B and C.

Table I shows that, when compared with Full SBS, not doing any search (No SBS) produces a big increase in standard error (from about .5 million to about 1.3 million). There is also an apparent problem with bias, most likely due to the quality of the data. Evidence from the 1990 PES evaluation program shows that some of the 1990 PES enumerators did not follow some instructions, specifically instructions related to the recording of the geographic location of housing units in the E sample. As a result, information on whether an E sample housing unit was found in the search area was not completely accurate. Plans A, B, and C involve selecting clusters with a large number of census geocode errors or household nonmatches with certainty and selecting a systematic random sample from among the remaining eligible clusters. Clusters with at least one census geocode error or one P-sample household nonmatch were included in the TES sampling universe. Other block clusters were excluded from the TES sampling universe.

The worst variance problem is with plan A, in which the 10% of clusters with the most TES-eligible housing units were included with certainty and the rest were sampled in. The problem with this plan turned out to be weighting. Some clusters have very high weights in the DSE, and when those clusters were sampled, the weights of the TES people in those clusters were multiplied by another factor of about 4. Plan B sought to correct that problem by selecting 5% of clusters with certainty based on the unweighted number of TES eligible housing units and another 5% based on the weighted number. This corrected much of the problem with increased variance. Plan C involves certainty selection based only on the weighted criterion. For the direct total DSE, it appears slightly better than Plan B, but for some subpopulation groups, the

Comparison of Certainty Only and Certainty with Sample TES Plans Direct Total DSE

	Total DSE	Sum SE's	Est. Coef. Var.
NO SBS	251,231,404	1,341,705	0.53%
Full SBS	247,530,864	479,219	0.19%
Plan A	247,766,573	507,560	0.20%
Plan B	247,711,502	492,840	0.20%
Plan C	247,860,084	486,497	0.20%



variances are larger than Plan B. We therefore are implementing Plan B for TES sampling.

LIMITATIONS

There were some issues related to TES which our 1990 simulation was not equipped to handle. One is the different methodology for handling movers planned for 2000. Our 1990 data files included match and correct enumeration probabilities for movers based on the methodology used in 1990 (Procedure B.) Trying to duplicate the 2000 methodology (Procedure C) on the 1990 data would have introduced numerous complexities into the analysis that would have made comparisons more difficult. We accept the change in the mover probabilities as a limitation on our results, and believe that it does not effect the comparisons between competing methods to an extent great enough to change our results.

Another potentially important change is greater emphasis on limiting the weights assigned to

clusters in the A.C.E. The fact that some clusters had very high weights was an important cause of the high variances shown by Plan A. It is likely that having lower A.C.E. cluster weights would significantly reduce the variances under Plan A. To simulate this effect as best we could, we replaced all cluster weights greater than 3,000 with that weight and estimated new variances under the condition of "capped weights." Five possible TES samples were generated. For each of the sample we calculated the DSE and the variance estimate of the DSE. The average DSE and standard error were calculated and are displayed in Table 1 below. The results were about what we expected – the variances were reduced under both, but the capped Plan A variances were still somewhat larger than those corresponding to Plan B. We therefore continue to recommend the use of Plan B for variance reduction reasons, even though the maximum 2000 weights should be lower than in 1990.

Table 1 - Five Runs of Plans A and B with original and capped Weights
(values in thousands)

	<u>DSE</u>	<u>Ave SE</u>	<u>Min SE</u>	<u>Max SE</u>
Plan A Original Weights	247,767	508	484	539
Capped Weights	247,883	497	481	515
Plan B Original Weights	247,712	493	490	495
Capped Weights	247,701	473	469	476

Table 2 - 1990 Block Clusters by Number of Geocoding Errors

Number of Geocode Errors	Number of Clusters	Count of HU	Weighted HU
115	1	115	133,345
111	1	111	29,247
84	1	84	43,954
67	2	134	45,533
66	1	66	17,871
61	1	61	128,814
46	1	46	16,024
42	1	42	5,359
35	1	35	9,990
33	1	33	139,962
30	1	30	16,157
29	1	29	2,077
28	1	28	5,018
26	1	26	87,847
25	2	50	6,670
23	1	23	273,525
21	2	42	39,604
20	6	120	83,594
19	3	57	22,547
18	4	72	37,864
17	5	85	70,831
16	5	80	46,873
15	4	60	43,390
14	1	14	14,885
13	5	65	50,273
12	9	108	60,326
11	11	121	130,281
10	6	60	34,138
9	9	81	66,122
8	11	88	55,731
7	17	119	77,999
6	20	120	67,613
5	19	95	64,581
4	40	160	136,978
3	57	171	128,476
2	123	246	152,758
1	325	325	197,593
0	4,480	0	0
Total	5,180	3,202	2,543,850

Table 3 - 1990 Block Clusters by Number of Household Non-Matches

Number of Non Matches	Number of Clusters	Count of HU	Weighted HU
85	1	85	129,328
77	1	77	41,240
71	1	71	107,946
67	1	67	2,034
46	1	46	41,225
40	1	40	5,558
38	2	76	17,411
37	2	74	53,325
36	1	36	3,466
34	1	34	12,627
33	1	33	14,476
32	1	32	12,386
31	2	62	23,192
30	2	60	105,031
29	2	58	83,451
28	1	28	43,635
27	3	81	168,141
26	2	52	93,399
25	7	175	114,178
24	4	96	54,874
23	1	23	8,017
22	6	132	45,787
21	3	63	140,203
20	6	120	41,589
19	8	152	114,804
18	8	144	86,387
17	9	153	96,427
16	12	192	131,421
15	13	195	81,553
14	18	252	148,641
13	10	130	62,337
12	20	240	59,104
11	19	209	90,314
10	20	200	134,211
9	44	396	271,967
8	62	496	273,366
7	88	616	491,102
6	100	600	532,405
5	147	735	414,736
4	243	972	608,183
3	410	1,230	724,996
2	629	1,278	821,558
1	1,219	1,219	734,029
0	2,038	0	0
Total	5,180	11,030	7,340,060

Figure 1

P sample Targeted Extended Search people

Status After HUFU:	Census Unedited File Status of Census Unit					
	No Change		Delete		Add	
	Hlds No Matches	Hlds ≥ 1 Matches	Hlds No Matches	Hlds ≥ 1 Matches	Hlds No Matches	Hlds ≥ 1 Matches
M, MU	Not TES	Not TES	Not TES	Not TES		
CI, UI	TES	Not TES			Not TES	Not TES
GI, ZI, DI	Removed from P Sample					

The shaded cells theoretically are empty.

Hlds - Households

Definition of codes for P sample housing units after Housing Unit Follow-Up (HUFU):

M= Matches to a Census address.

MU= Matches to a Census address but not with certainty.

CI= Does NOT match to a Census address. The P sample housing unit is in the block cluster.

UI= Does NOT match to a Census address but the P sample housing unit is NOT confirmed to be in the block cluster.

GI= Exists but NOT in the block cluster (geocoding error).

ZI= Does NOT exist in the block cluster (unit burned down).

DI= Should NOT have been listed (duplicate)..

Figure 2

E sample Targeted Extended Search People

Status After HUFU:	TES FU	Census Unedited File Status of Census Unit					
		No Change		Delete		Add	
		Hlds No Matches	Hlds ≥ 1 Matches	Hlds No Matches	Hlds ≥ 1 Matches	Hlds No Matches	Hlds ≥ 1 Matches
M, CE, UE, EE, DE		Not TES	Not TES	Not TES	Not TES		
GE	GS GC GE GU	TES	Not TES	Not TES	Not TES		
						Not TES	Not TES

The shaded cells theoretically are empty.
Hlds - Households

Definition of codes for E sample housing units after Housing Unit Follow-Up (HUFU):

- M= Matches to a P Sample address.
- CE= Exists in the block cluster but is NOT in the P sample.
- UE= Does NOT match to a P sample address but the Census housing unit is NOT confirmed to be in the block cluster.
- EE= Does NOT exist within the block cluster.
- DE= Is erroneously enumerated (duplicate).
- GE= Exists but NOT in the block cluster (geocoding error).



UNITED STATES DEPARTMENT OF COMMERCE
Bureau of the Census
Washington, DC 20233-0001

January 13, 2000

MASTER FILE

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-19

MEMORANDUM FOR: Howard Hogan
Chief, Decennial Statistical Studies Division

From: Donna Kostanich *DK*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared By: Patrick Cantwell *PC*
Statistical Communications

Subject: Accuracy and Coverage Evaluation Survey: Missing Data
Procedures

1. Introduction

This document gives a general overview of plans for handling missing data in the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) as well as details of some of these procedures. It provides information and detail beyond that given in the DSSD Procedures and Operations Memorandum Series #Q-3. (See references.)

Section 2 contains general background and describes missing data procedures used in past censuses and tests. Section 3 provides some details of the Census 2000 procedures. A noninterview adjustment procedure, outlined in Section 3.1, is used to account for whole-household nonresponse. A characteristic imputation procedure, outlined in 3.2, is used to assign values for specific missing demographic variables. Finally, persons with unresolved match, residence, or enumeration status have probabilities assigned according to a procedure outlined in 3.3.

The missing data procedures to be used in the 2000 Census are similar to those used on the Integrated Coverage Measurement (ICM) sample in the Census 2000 Dress Rehearsal. An outline of the ICM procedures and a summary of related research is given in Ikeda, Kearney, and Petroni (1998). An overview of the results for missing data in the Dress Rehearsal is given in Kearney and Ikeda (1999).

2. General Background

Census 2000 will be conducted for the entire nation. There are two separate Accuracy and Coverage Evaluation (A.C.E.) samples: one for the 50 states and the District of Columbia, and a second sample for Puerto Rico. Although the National and Puerto Rico samples will be processed separately by the A.C.E. missing data system, the procedures will be similar for the two. For simplicity, this document will address only the National sample.

The A.C.E. will rely on dual system estimation (DSE) to determine estimates. The Census Bureau obtains a roster from the A.C.E. blocks independently of the Census. The independent roster (P Sample) and the Census roster (E Sample) will be matched; the results of the matching will then be used to estimate the number of persons missed by both rosters. Estimates are calculated separately within population subgroups called post-strata. Post-stratum estimates are then used to determine adjustment factors to be applied to all people counted in the Census according to their specific post-stratum. Finally, adjusted counts for any geographic area will be calculated by summing the adjusted counts of people in the area. An appropriate rounding method is applied to produce integer counts of people at all levels.

The formula for dual system estimation is as follows:

$$D\hat{S}E = (C - II) \left(\frac{CE}{N_e} \right) \left(\frac{N_p}{M} \right)$$

where

C = Census total records, including imputed, duplicate, fictitious etc. (the Census count),

II = number of whole-person census imputations,

CE = weighted estimate of appropriate, unique, complete and correct enumerations,

N_e = weighted E-Sample total, including duplicate, fictitious, etc.

N_p = weighted P-Sample total,

M = the estimated number of persons from the P Sample population who match to the E Sample population

Note: Persons in Group quarters are excluded from the A.C.E. and, thus, from the above counts.

There are two main types of missing data in the A.C.E. and three processes used to correct for them. It should be noted that the term missing data applies after all follow-up attempts have been made. The first type is *unit missing data*. These are households that were not interviewed in the A.C.E. either because they could not be contacted or because the interview was refused. The noninterview adjustment process spreads the weights of these households among households that were interviewed in the same noninterview cell.

The other type of missing data is *item missing data*. This situation occurs when we have some

information for a household or person but portions of the data are missing. We consider two groups of item missing data: demographic items and items relating to a specific status. Certain demographic items are imputed because they are necessary to assign people to post-strata. Some demographic missing data items, such as race, are imputed using a hot-deck procedure (for the most part). Others, such as age, are imputed based on available demographic distributions.

For a small number of people in the P Sample, there is not enough information available to determine the match status (whether or not the person matches to someone in the E Sample in the same block cluster) or the residence status (whether or not the person was living in the block cluster on Census Day). Determining residence status is important for P-Sample people because only Census-Day residents contribute toward N_p , the size of the P Sample, and M , the estimated number of matches in the P Sample. Similarly, for people in the E Sample, there may not be enough information to determine whether the person was correctly enumerated. Such cases where status cannot be determined are said to be "unresolved." Generally for cases with missing status a probability is assigned based on information available about the specific case and about cases with similar characteristics.

There is some debate on how to assign the probabilities of match or residence (P Sample) or enumeration (E Sample). The Bureau's missing data team has considered several options and examined the available research on different methods of imputation. The currently planned procedure is described in Section 3.3.

In the 1990 Post-Enumeration Survey, a hierarchical logistic regression program was used to calculate missing probabilities of match and correct enumeration. (Due to the procedure used to treat movers in 1990, unresolved residence status was not a concern then.) The model and some results are discussed in Belin et al. (1993). Research using data from the census tests in 1995 and 1996 as well as the Dress Rehearsal in 1998 indicates that the exact method of calculating probabilities for unresolved status (match, residence, or correct enumeration) has a limited effect on the dual system estimates. More details of this research can be found in memo #Q-3. Other references are listed at the end of this document.

Several research projects are currently in progress within the missing data team. These include using different variables to form the imputation cells, comparing imputation cell estimation to logistic regression modeling, and evaluating the sensitivity of the DSE to various missing data treatments. We plan to use imputation cell estimation to determine these probabilities unless the research in progress suggests we should do otherwise.

3. Missing Data Methodology

Following is a summary of the procedures planned to address missing data in the 2000 Census Accuracy and Coverage Evaluation. As was mentioned in the previous section, some changes are under consideration subject to the results of research.

3.1 Household Noninterview Adjustment

Before describing the noninterview adjustment procedure, we first describe the manner in which movers are handled. Census 2000 will use a procedure called Procedure C to handle cases in which a person has moved between Census Day and the day of the A.C.E. interview. (See definitions below.)

non-mover: a person who lives in the sample housing unit on Census Day *and* on the day of the A.C.E. interview

in-mover: a person who lives in the sample housing unit on the day of the A.C.E. interview, but lived elsewhere on Census Day

out-mover: a person who lived in the sample housing unit on Census Day, but has moved elsewhere before the A.C.E. interview.

For example, if a mother decided to move in with her grown son after Census Day but before the A.C.E. interview, then she would be considered an in-mover while her son would be considered a non-mover.

Procedure C uses in-movers (along with non-movers) to estimate the number of P-Sample people in the post-stratum, while using out-movers to estimate the match rate of movers in the post-stratum. Mover status is assigned at the time of the CAPI interview. Questions are asked to determine who currently lives in the household and who lived in the household on Census Day. Thus two rosters are created for each household, the Census Day roster and the Interview Day roster.

Noninterview adjustment is only performed on the P-Sample. The procedure is similar to that used in the Census 2000 Dress Rehearsal. Because of the use of Procedure C estimation, there are two noninterview adjustments--one based on housing-unit status as of Census Day (i.e., the Census Day roster), and the other based on housing-unit status as of the day of the A.C.E. interview (i.e., the Interview Day roster). For occupied housing units, the definitions of interview and noninterview are as follows:

interview: A unit is an interview (for the given reference date--Census Day or Interview Day) if there is at least one person (with name and at least two demographic characteristics) who possibly or definitely was a resident of the housing unit on the given reference date.

noninterview: An occupied housing unit (as of the given reference date) that is not an interview is a noninterview.

Note that some housing units are vacant, and do not contribute toward the noninterview

adjustment or dual system estimation. Consider the same example as above--the mother moving in with her son. In this case the household would be an interview for both Census Day and Interview Day since both rosters would have at least one person for whom we have sufficient information.

However, consider a case in which a family moves into an apartment just after Census Day and has no knowledge of the previous residents. Suppose also that the interviewer is unable to find information about the previous residents from the landlord or any other source. When the A.C.E. interview is conducted the Interview Day roster will be complete but the Census Day roster will be empty since we are unable to obtain any information about the Census Day residents. Thus the housing unit would be considered an A.C.E. interview but a Census Day noninterview.

If we find that a housing unit was vacant on Census Day then that household does not factor into the Census Day noninterview adjustment. Similarly housing units that were vacant at the time of the A.C.E. interview do not factor into the Interview Day noninterview adjustment.

Each of the two noninterview adjustments spreads the weights of noninterviewed units over interviewed units in the same noninterview cell, as defined by the block cluster and the type of basic address. For purposes of this adjustment, the type of basic address is grouped by single-family, apartments, and other. The Census Day housing unit status is used across all P-Sample units in the Census Day noninterview adjustment, which is then used to adjust the person weights of non-movers and out-movers. Similarly, A.C.E. Interview Day housing unit status is used in the Interview Day noninterview adjustment, which is then used to adjust the person weights of in-movers. The formulae are as follows:

For a given cluster and type of basic address, the Census Day noninterview adjustment factor for the j^{th} cluster is computed as

$$f^{*}_{c,j} = \frac{\sum_{\text{Census Day interviews}} w_i + \sum_{\text{Census Day noninterviews}} w_i}{\sum_{\text{Census Day interviews}} w_i}$$

For a given cluster and type of basic address, the Interview Day noninterview adjustment factor for the j^{th} cluster is computed as

$$f^*_{a,j} = \frac{\sum_{\substack{\text{Interview Day} \\ \text{interviews}}} w_i + \sum_{\substack{\text{Interview Day} \\ \text{noninterviews}}} w_i}{\sum_{\substack{\text{Interview Day} \\ \text{interviews}}} w_i}$$

where w_i represents the weight of housing unit i , the inverse of its probability of selection into the A.C.E. sample. This weight does not include any sampling for targeted extended search of the block cluster.

If the number of interviewed units (in a given block cluster and type of basic address category) is too small compared to the number of noninterviewed units, then the weights of the noninterviewed units are spread out over a broader category of interviewed units. There are collapsing rules to determine which broader category to use.

3.2 Characteristic Imputation

When missing, the characteristics race, Hispanic origin, sex, tenure, and age will be imputed in the A.C.E. Because only these variables are needed to assign sample people to post-strata for dual system estimation, characteristic imputation is not carried out for other missing variables (with the exception of the unresolved status items discussed in Section 3.3).

For the Census 2000 E Sample we use demographic information from the Census 2000 Census edited file (CEF). Therefore the only A.C.E. imputation that would need to be done in the E-Sample is for E-Sample persons that cannot be matched to the CEF. In the Dress Rehearsal, all E-Sample persons matched to the CEF; we expect the same for the Census 2000 E-Sample. The methodology for any remaining E-Sample A.C.E. imputation is essentially the same as that for the P-Sample.

P-Sample characteristic imputation for Census 2000 is nearly identical to characteristic imputation for the Dress Rehearsal A.C.E.. Some demographic missing data items, such as race, are imputed generally using a hot deck procedure. Others, such as age, are imputed based on the available demographic distributions. P-Sample person mover status is not considered when imputing characteristics. (However, for purposes of imputation, a P-Sample whole household of in-movers on Interview Day is considered to be a separate household from the P-Sample whole household of out-movers living in that housing unit on Census Day.)

Imputation for a specific missing characteristic is not affected by the imputation for other missing characteristics. Before imputation begins, age and sex distributions are calculated nationally using the P-Sample data. For hot-deck imputation, the data are sorted by cluster, then map spot number, then unit identifier. This essentially produces a geographic sort.

Tenure. Tenure is imputed from the previous household with a similar type of basic address with

tenure recorded.

Race. Generally missing race is imputed from the distribution of race in the same household. If everyone in the household is missing the variable race, then the distribution of the nearest previous household with reported race and the same recoded Hispanic origin is used. The flowchart in Attachment 1 illustrates the procedure. All 63 possible combinations of the 6 basic race categories are imputed (the 6 basic categories being White, Black, American Indian or Alaskan Native, Asian, Native Hawaiian or Other Pacific Islander, Other). All 63 categories are treated the same in the imputation (that is, there aren't any special procedures for any categories or groups of categories).

Hispanic origin. Hispanic origin is imputed in a manner analogous (and symmetric) to that for imputing missing race.

Age. For most relationship-to-reference-person categories in multi-person households, age is imputed from the distribution of age for persons with similar relationship to reference person, and similar age of reference person. For the collapsed other-relative category and the collapsed nonrelatives category, age is imputed from the distribution of age for persons with a similar relationship category in multi-person households. If relationship is missing we impute from the distribution of age in multi-person households. For one-person households, age is imputed from the distribution of age in one-person households. See Attachment 2.

Sex. Sex of reference person (with spouse present) or spouse of reference person is imputed by assigning the person with a missing value for sex the sex opposite to that of their spouse. If both reference person and spouse have sex missing, then sex for the reference person is imputed from the distribution of sex for reference persons with spouse present. The spouse is then assigned the sex opposite to that of the reference person. For one-person households, sex is imputed from the distribution of sex in one-person households. For the reference person (with no spouse present) of a multi-person household, the distribution of sex for reference persons of multi-person households with no spouse present is used. For persons (except reference persons and spouses) from multi-person households with non-missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons and spouses) from multi-person households. For persons from multi-person households with missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons) from multi-person households. See Attachment 3.

3.3 Match, Residence, and Correct Enumeration Probabilities for Unresolved Cases

We will use imputation cell estimation to assign probabilities for P-Sample people with unresolved match or residence status, and for E-Sample people with unresolved enumeration status. P- (or E-)Sample persons are separated into groups called imputation cells based on operational, demographic, or geographic characteristics. We are also considering other characteristics to define the imputation cells. The weighted average of 1's and 0's (representing,

e.g., match and non-match, respectively) is calculated for each cell, and that value is imputed for all persons in the cell with missing probabilities.

We note that the noninterview adjustment factor is *not* incorporated when we calculate these averages. The reason for this is that the noninterview adjustment is applied to spread the weight of noninterviewed housing units over interviewed units. However, all persons in noninterviewed units will be either nonresidents or unresolved (since, by definition, if one person in the household is a resident then the household is considered an interview). Therefore, using the noninterview factor to calculate the averages for unresolved cases would produce a biased estimate of residence probability. Thus we use the initial weights here. The issue is moot when resolving E-Sample cases with missing enumeration status, as a noninterview adjustment is not applied to E-Sample persons. (See Section 3.1.)

Match Status. Each Census Day resident or possible resident j in the P Sample has some probability, $Pr_{m,j}$, of matching to a person in the E Sample. This probability equals 1 if the person matches, and 0 if the person does not match. Persons whose match status is unknown or unclear must be assigned a match probability between 0 and 1. We assign this probability based on all confirmed or possible Census-Day residents in the P-Sample that have a resolved match status.

For people with unresolved match status, we estimate the match probability. Separate averages are calculated for person non-movers and person out-movers (other variables are being considered). The match probability for persons with unresolved match status is the weighted proportion of matches in the same mover category among persons with resolved final match status (excluding confirmed Census Day nonresidents).

$$Pr_{m,j} = \begin{cases} 1 & \text{if person } j \text{ is a match on Census Day} \\ 0 & \text{if person } j \text{ is NOT a match on Census Day} \\ Pr^*_{m,j} & \text{if person } j \text{ is unresolved} \end{cases}$$

For each imputation cell, the estimated match probability is:

$$Pr^*_{m,j} = \frac{\sum_{\text{resolved units}} w_i Pr_{m,i}}{\sum_{\text{resolved units}} w_i}$$

Most persons with unresolved match status are persons with insufficient information for matching. Person in-movers are not sent to matching because they were not residents on Census Day. Thus the match probability for person in-movers is irrelevant to estimation and is set to 0.

Residence Status. Similar to match status, each P-Sample person j has some probability, $Pr_{res,j}$, of being a resident in the sampled block at the time of the census. For those whose residence status is unknown, this probability has to be estimated. For this purpose we use the P-Sample people with a resolved final residence status and a final match code status (i.e., all who went into the person matching operation). We separate them by match code group and calculate a weighted ratio for each group. Seven P-Sample match code groups are used:

- 1 = matches needing follow-up,
- 2 = possible matches,
- 3 = nonmatches needing follow-up from partial household nonmatches,
- 4 = nonmatches needing follow-up from whole-household nonmatches,
- 5 = nonmatches needing follow-up from conflicting households,
- 6 = persons resolved before follow-up, and
- 7 = persons with insufficient information for matching.

The details of these codes can be found in Childers (2000).

The residence probability estimated for unresolved persons in match code groups 1 through 6 is the weighted proportion of persons (among cases with resolved residence status) in the given match code group who are residents. The residence probability for persons with insufficient data for matching (group 7) is the weighted proportion of all persons who are residents.

$$Pr_{res,j} = \begin{cases} 1 & \text{if person } j \text{ is a resident on Census Day} \\ 0 & \text{if person } j \text{ is NOT a resident on Census Day} \\ Pr^*_{res,j} & \text{if person } j \text{ is unresolved} \end{cases}$$

For each imputation cell, the estimated residence probability is:

$$Pr^*_{res,j} = \frac{\sum_{resolved\ units} w_i Pr_{res,i}}{\sum_{resolved\ units} w_i}$$

The proportions are based jointly on person non-movers and person out-movers with resolved

final residence status. The Census Day residence probability for person in-movers is irrelevant to estimation and is set to 0. Note that the residence probability as of the date of A.C.E. interview for person in-movers and person non-movers is assumed to be 1 (except that infants born after Census Day are not considered to be A.C.E. interview day residents).

Correct Enumeration Status. In the E Sample, each person j has some probability, $Pr_{ce,j}$, of having been correctly enumerated in the census. A key factor in determining this probability is the E-Sample person's final match code. Cases that remain unresolved will have their probability of correct enumeration imputed. Here we base the probability on the set of E-Sample people with resolved enumeration status (before accounting for duplication with non-E-Sample people).

For E-Sample persons with unresolved enumeration status, the correct enumeration probability is the weighted proportion of correct enumerations (among persons with resolved enumeration status) in the given match code group.

$$Pr_{ce,j} = \begin{cases} 1 & \text{if person } j \text{ is correctly enumerated} \\ 0 & \text{if person } j \text{ is NOT correctly enumerated} \\ Pr_{ce,j}^* & \text{if person is unresolved} \end{cases}$$

For each imputation cell, the estimated probability of correct enumeration is

$$Pr_{ce,j}^* = \frac{\sum w_i Pr_{ce,i}}{\sum_{\text{resolved units}} w_i}$$

There are nine E-Sample match code groups:

- 1 = matches needing followup,
- 2 = possible matches,
- 3 = nonmatches from partial household nonmatches,
- 4 = nonmatches from whole-household nonmatches where the housing unit matched in housing unit matching,
- 5 = nonmatches from conflicting households where the E-Sample housing unit was not in regular nonresponse followup,
- 6 = nonmatches from conflicting households where the E-Sample HU was in regular nonresponse followup,

7 = nonmatches from whole-household nonmatches where the housing unit did not match during housing unit matching,

8 = persons resolved before followup, and

9 = persons with insufficient information for matching.

The details of these codes can be found in Childers (2000).

There is an additional adjustment made due to (i) duplication with persons subsampled out of the E-Sample (in large clusters), and (ii) duplication with Census group quarters persons in the same A.C.E. block cluster. If an E-Sample person is duplicated with k persons who are either subsampled out of the E-Sample or from Census group quarters, then the initial correct enumeration probability is multiplied by $1/(k + 1)$, since we do not know which person is the "real" person.

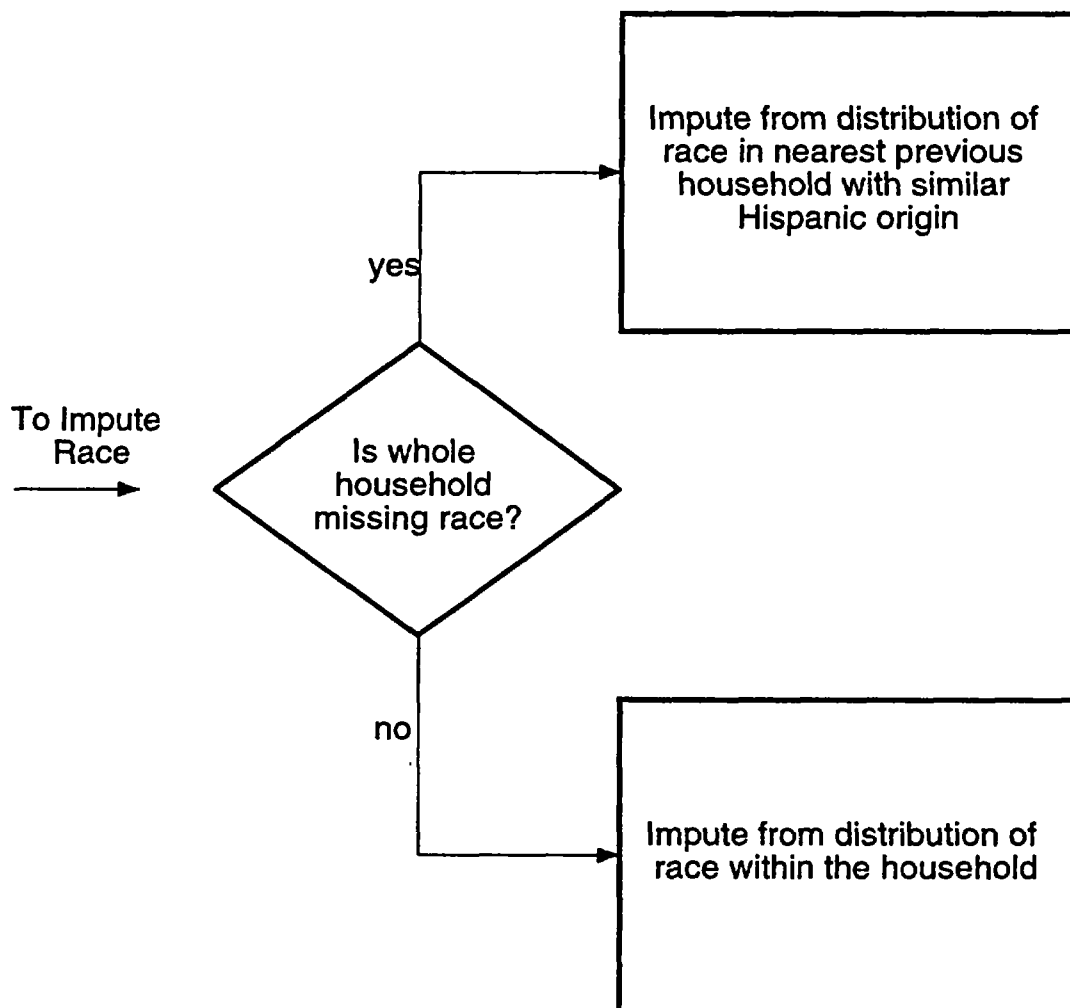
References

- Belin, T., G. Diffendal, S. Mack, D. Rubin, J. Schafer, and A. Zaslavsky (1993). "Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation," *Journal of the American Statistical Association*, 88, pp. 1149-1166.
- Cantwell, P. (1999). "Accuracy and Coverage Evaluation Survey: Overview of Missing Data for P & E Samples," DSSD Census 2000 Procedures and Operations Memorandum Series #Q-3.
- Childers, D. (2000). "The Design of the Census 2000 Accuracy and Coverage Evaluation," DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1.
- Ikeda, M., A. Kearney, and R. Petroni (1998). "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," *Proceedings of the Survey Research Section, American Statistical Association*, pp. 617-622.
- Ikeda, M. (1997). "Effect of Using the 1996 ICM Characteristic Imputation and Probability Modeling Methodology on the 1995 ICM P and E-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series A-20.
- Ikeda, M. (1997). "Effect of Different Methods for Calculating Match and Residence Probabilities for the 1995 P and E-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series A-23.
- Ikeda, M. (1997). "Effect of Different Methods for Calculating Correct Enumeration Probabilities for the 1995 E-Sample Data," DSSD Census 2000 Dress Rehearsal Memorandum Series A-28.

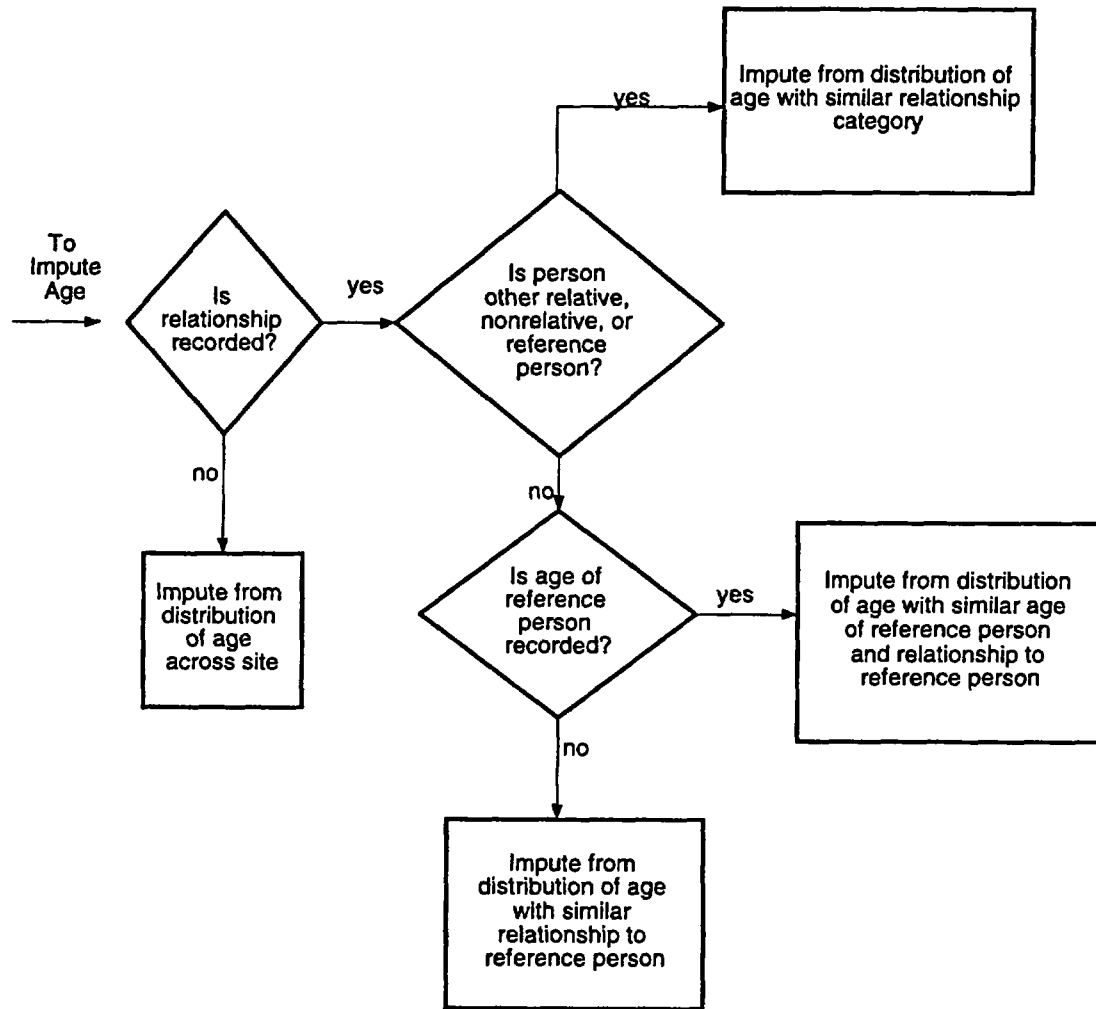
Ikeda, M. (1997). "Effect of Using Simple Ratio Methods to Calculate P-Sample Residence Probabilities and E-Sample Correct Enumeration Probabilities for the 1995 Data," DSSD Census 2000 Dress Rehearsal Memorandum Series A-30.

Kearney, A. and M. Ikeda (1999). "Handling of Missing Data in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample," *Proceedings of the Survey Research Section, American Statistical Association*, to appear.

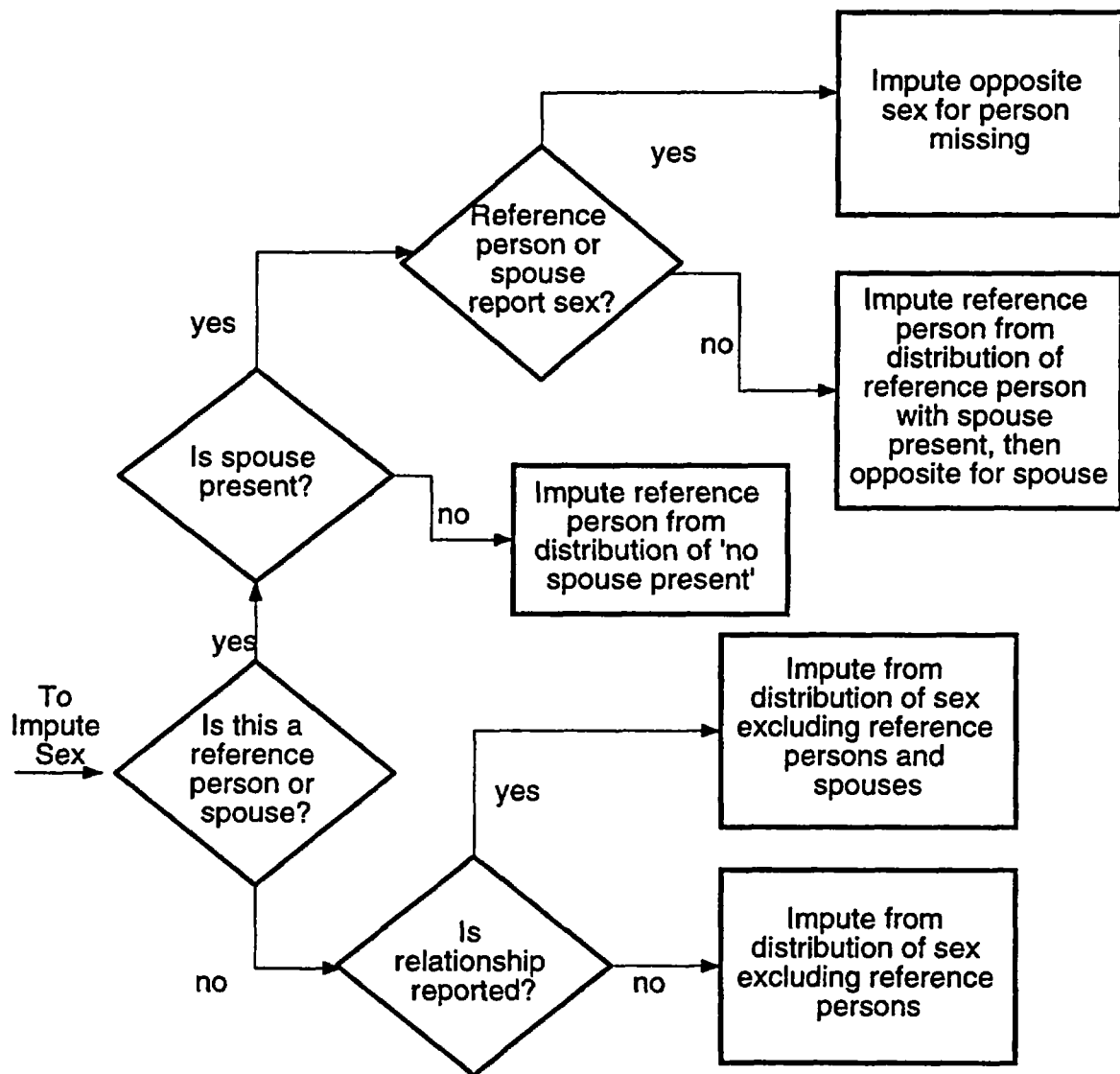
Imputation of Race for the P Sample



Imputation of Age for the P Sample



Imputation of Sex for the P Sample





MASTER FILE

January 12, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q -20

MEMORANDUM FOR Howard Hogan
 Chief, Decennial Statistical Studies Division

From: Donna Kostanich *DK*
 Assistant Division Chief, Sampling and Estimation

Prepared by: Richard Griffin *R.G.*
 Estimation Team

Subject: Accuracy and Coverage Evaluation Survey: Dual System
 Estimation

1. INTRODUCTION

Dual System Estimation (DSE) within post-strata is planned for the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) Survey. DSE was used by the Census Bureau to estimate Census Coverage for the 1980 Census Post Enumeration Program (PEP) and the 1990 Census Post Enumeration Survey (PES). The use of DSE in the 1980 PEP is described in Fay (1988), while Hogan (1992,1993) describes the use of DSE in the 1990 PES. As described in Killion (1998), several alternatives to DSE were considered for Census 2000. These included CensusPlus and the use of logistic regression estimated inclusion probabilities to correct for heterogeneity bias. CensusPlus was tested in the 1995 and 1996 Census tests and was shown to produce grossly inferior results to DSE. The research for the logistic regression approach did not provide conclusive findings to date. Research on the logistic regression approach will continue as an evaluation tool for Census 2000 and for 2010 Census planning.

Section 2 describes the basic DSE model including a discussion of the need for post-stratification. Section 3 presents the detailed DSE that will be computed within each final post-stratum for Census 2000. All components of the DSE are defined. The DSE includes special handling of missing data, search area for matching, and movers. Missing data and search area for matching are covered in detail in Cantwell (2000) and Navarro (2000) respectively. Section 3 will provide a brief overview of these topics. The method we are using for dealing with movers in Census 2000 DSE is also discussed in Section 3 while the Appendix provides

detailed background on options for dealing with movers in census coverage measurement surveys as well as a derivation comparing the correlation bias of these options. Finally, Section 4 concludes the paper with a short discussion of how the results from DSE serve as input to synthetic estimation at the block level. A detailed discussion of synthetic estimation is provided in Haines (2000).

2. DSE MODEL AND POST-STRATIFICATION

The DSE model is discussed in detail in Wolter (1986) and more generally in Hogan (1992). This paper gives a general presentation.

The DSE model (applied within each post-stratum) conceptualizes each person as either in or not in the Census enumeration, as well as either in or not in the A.C.E..

	In Census	Out of Census	Total
In A.C.E.	N_{11}	N_{12}	N_{1+}
Out of A.C.E.	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

All cells are conceptually observable except N_{22} , and any of the marginal cells that include N_{22} . The model assumes independence between the Census and the A.C.E.. This means that the probability of being in the ij th cell, p_{ij} , is the product of the marginal probabilities, $p_{i+}p_{+j}$ is the estimate of total population in a post-stratum with the independence assumption.

$$DSE = N_{++} = \frac{(N_{+1})(N_{1+})}{N_{11}}$$

The independence assumption can fail either due to causal dependence between the Census enumeration and the A.C.E. enumeration or due to heterogeneity in capture probabilities within the post-stratum. Causal dependence occurs when the event of an individuals inclusion or exclusion from one system affects their probability of inclusion in the other system. For example, some people who did answer the census may not cooperate with the A.C.E., thinking they had helped enough. As another example, a person contacted during A.C.E. listing may not respond to the Census thinking that the A.C.E. Lister already recorded them. However, even if causal independence is true for all individuals ($p_{ij} = p_{i+}p_{+j}$), the independence assumption can be

violated by heterogeneity. Either the Census inclusion probabilities p_{+1} OR the A.C.E. inclusion probabilities p_{1+} must be equal for all individuals (homogeneity in both systems is not required). Failure of the independence assumption for either reason results in correlation bias.

Post-stratification or grouping of individuals likely to have similar inclusion probabilities and calculating DSEs within post-strata is done to decrease correlation bias. Details of the Census 2000 post-stratification research methodology are given in Kostanich (1999) and Griffin (1999). Results of this research and the post-stratification design chosen for Census 2000 A.C.E. are given in Griffin (2000).

The DSE can be written as follows:

$$DSE = N_{+1} \left(\frac{N_{1+}}{N_{11}} \right)$$

That is, the total population estimate is estimated by the number captured in the Census times the ratio of those captured in the A.C.E. survey to those captured in both systems.

In practice the components of the DSE are estimated from a sample survey. N_{+1} is not the Census count; the census count (C) must be corrected for erroneous enumerations as well as for persons enumerated in the Census with insufficient information to match to the A.C.E. enumeration. To actually estimate the number of people correctly enumerated in the Census, a sample of all data defined persons is selected. This sample of data defined census persons is called the enumeration or E sample. To estimate the ratio of those captured in both systems to those captured in A.C.E., the population or P sample is used. The P sample consists of persons interviewed during A.C.E. enumeration.

$$D\hat{S}E = (C - II) \left(\frac{CE}{N_e} \right) \left(\frac{N_p}{M} \right)$$

The form of the DSE used in Census coverage measurement surveys such as A.C.E. is as follows:

where,

C = the census count

II = the number of census enumerations that are non-data defined persons or imputed persons

CE = the estimated number of correct enumerations from the E sample (persons in the E sample without sufficient information for matching are NOT classified as correct enumerations)

N_e = the estimated total population from the E sample

N_p = the estimated total population from the P sample

M = the estimated number of persons from the P sample population who match to the E sample population

Note: Persons in Group quarters are excluded from all the above counts.

3. Dual System Estimation

3.1 Definitions

Block Cluster: A grouping of one or more census blocks. Block clusters are the primary sampling units for A.C.E. and average about 30 housing units each.

Correct Enumeration (CE): A correct enumeration is a person who is enumerated in a block cluster during the census who is also determined by A.C.E. operations to have lived in that block cluster (or if appropriate a surrounding block) on Census Day. Correct enumerations have a correct enumeration probability, $Pr_{ce,j}$, equal to 1 for each person j .

Correct Enumeration Probability ($Pr_{ce,j}$): This is defined as the probability that person j in the E Sample was correctly enumerated in the A.C.E. (or surrounding block) block cluster. The probability of correct enumeration is typically 0.00 or 1.00 but it can take on values within this range due to missing data imputation.

Coverage Correction Factor (CCF): The coverage correction factor for a post-stratum is calculated by dividing the DSE for that post-stratum by its census count. A.C.E. estimates for any population at any level are obtained by multiplying the coverage factor by the census count within each post-stratum, then summing over all post-strata.

Data-Defined Person: This concept is defined for all census persons. A data-defined person is a person who has two or more of the 100% data items answered on their census form. Any items can be selected from the 100% data items, which include name, age, sex, race, and Hispanic origin. Relationship to person 1 is also a 100% data item for all persons besides person 1. Persons not satisfying this criteria are referred to as non-data defined.

E Sample: The E Sample is the Enumeration Sample. It consists of all data-defined persons in the A.C.E. block clusters who were enumerated in the census.

Group Quarters (GQ) Counts: The number of persons living in GQs, such as college dormitories, prisons, or military barracks. GQ persons are excluded from the A.C.E. universe.

Inmover: A person who has moved into an housing unit in an A.C.E. block cluster after Census Day.

Insufficient Information in Census (II): Those persons in the census for whom there is insufficient information for inclusion in the E Sample. Very little data is available for these persons. This category includes non-data-defined persons and persons in whole household imputations. Note that insufficient information in Census is different than insufficient information for matching. The former are excluded from the E-sample and the latter are in the E sample.

Match Probability (Pr_{mj}): This is defined as the probability that person j in the P Sample was matched to a person in the E Sample (or in a surrounding block). The match probability is typically 0.00 or 1.00 but it can take on values within this range due to missing data imputation.

Mover Status: Each person in the P Sample will be classified as a nonmover, outmover, or inmover. Nonmovers are persons who remain in the same housing unit between Census Day and A.C.E. interview day. Outmovers are persons who moved out of an housing unit in an A.C.E. block cluster between Census Day and A.C.E. interview day. Inmovers are persons who moved into an housing unit in an A.C.E. block cluster between Census Day and A.C.E. interview day.

Nonmover: An A.C.E. sample person whose housing unit on Census Day and the A.C.E. interview day are identical.

Outmover: A person who has moved away from an housing unit in an A.C.E. block cluster since Census Day.

P Sample: Also known as the Person Sample. The P Sample consists of those persons confirmed to be residents of the housing units in the A.C.E. block clusters as of Census Day by the independent portion of the A.C.E. reinterview and subsequent operations.

3.2 DSE Formula

The DSE for any given post-stratum is defined as (all counts and estimates are for a specific post-stratum):

$$D\hat{S}E = (C - II) \left(\frac{CE}{N_e} \right) \left[\frac{N_n + N_i}{\left(M_n + \left(\frac{M_o}{N_o} \right) N_i \right)} \right]$$

Adjustments to this DSE are made to handle the unlikely event that the formula results in division by zero.

The coverage correction factor for each post-stratum is defined as $CCF = \frac{D\hat{S}E}{C}$

The census count (unweighted) of persons in the post-stratum is C. It is possible (although not planned) that there will be late census data (added and deleted housing units) not included in data files used to obtain C for the DSE calculation. If this occurs, C will include persons in late deletes and exclude persons in late adds. For adds, lower census counts in the DSE (lowers the DSE) will be compensated by a lower match rate (raises the DSE) due to the persons in late adds not being included in the E sample. For deletes, larger census counts in the DSE (raises the DSE) will be compensated by a lower correct enumeration rate (lowers the DSE) due to persons in late deletes being included in the E sample. If there is late census data the denominator of CCF will be the census count from the final census file which includes persons in adds and excludes persons in deletes. This is necessary since the coverage factors are applied to the final census file.

II is the number of census people in the post-stratum with insufficient information for inclusion in the E Sample. This includes non-data-defined persons as well as whole household imputations. This count is based on unweighted census data.

The estimated # of E-sample persons is:

$$N_e = \sum_{j \in E\text{-Sample}} W_j$$

where, W_j = inverse of probability of selection, including a factor for Targeted Extended Search sampling.

The estimated # of “correct” enumerations is:

$$CE = \sum_{j \in E\text{-Sample}} Pr_{cej} W_j$$

Where, Pr_{cej} is:

- 1 if person j “correctly” enumerated
- 0 if person j NOT “correctly” enumerated
- Pr_{cej}^* if person j is unresolved
- Pr_{cej}^* is estimated through missing data imputation

Some persons will move between Census Day and A.C.E. interview day. A mover is a person whose location on the day of the A.C.E. interview differs from their location on Census Day. The treatment of movers has important ramifications for estimation. For Census 2000, movers are being treated by a procedure known as Procedure C (See Appendix). This procedure identifies all current residents living or staying at the sample address at the time of the A.C.E. interview, plus all other persons who lived at the sample address on Census Day who have since moved. For outmovers, the interviewers will attempt a proxy interview to obtain data such as name, sex, and age that can be used for matching. The match rate of movers is obtained using outmovers. On the other hand, the number of movers in the P Sample for A.C.E. sample areas is estimated by the in-movers. Note that no matching is done for in-movers.

N_n is the weighted total population for nonmovers for the post-stratum from the P sample. The weight for each person j is the product of three values

- (1) the inverse of the P sample selection probability including a factor for the Targeted Extended Search sampling (W_j),
- (2) a noninterview adjustment based on Census Day interview status ($f_{c,j}^*$), and
- (3) a Census Day residence probability ($Pr_{res,j}$)

The estimated # of P-sample nonmovers is:

$$N_n = \sum_{j \in \text{Nonmovers}} f_{c,j}^* Pr_{res,j} W_j$$

where, $Pr_{res,j}$ is:

- 1 if person j is a resident on Census Day
- 0 if person j is NOT a resident on Census Day

$Pr_{res,j}^*$ if person j is unresolved
 $Pr_{res,j}^*$ is estimated through missing data imputation

The estimated # of P-sample nonmover matches is:

$$M_n = \sum_{j \in \text{Nonmovers}} Pr_{m,j} f_{c,j}^* Pr_{res,j} W_j$$

where, $Pr_{m,j}$ is:

1 if person j is a match on Census Day
 0 if person j is NOT a match on Census Day
 $Pr_{m,j}^*$ if person j is unresolved
 $Pr_{m,j}^*$ is estimated through missing data imputation

N_i is the weighted total population for inmovers for the post-stratum from the P sample. The weight for each person j is the product of two values

- (1) the inverse of the P sample probability of selection (W_j) and
- (2) a noninterview adjustment factor based on A.C.E. interview day status ($f_{a,j}^*$)

The estimated # of P-Sample Inmovers is:

$$N_i = \sum_{j \in \text{inmovers}} f_{a,j}^* W_j$$

Note that all inmovers are assumed to be A.C.E. interview day residents.

The estimated # of P-Sample Outmovers is:

$$N_o = \sum_{j \in \text{Outmovers}} f_{c,j}^* Pr_{res,j} W_j$$

The estimated number of P-Sample Outmover Matches is:

$$M_o = \sum_{j \in \text{Outmovers}} Pr_{m,j} f_{c,j}^* Pr_{res,j} W_j$$

3.3 Missing Data

Cantwell (2000) provides a detailed discussion on the application of procedures for handling missing data for the DSE. There are three missing data procedures: household noninterview adjustment, characteristic imputation, and the assignment of missing match, residence and correct enumeration probabilities.

3.3.1 Household Noninterview Adjustment

Noninterview adjustment is only performed on the P sample. There are two noninterview adjustments which are basically identical to each other except for the reference date. One noninterview adjustment is based on housing unit status as of Census Day and is used to adjust the sampling weights of nonmovers and outmovers. The other noninterview adjustment is based on housing unit status as of the day of the A.C.E. interview and is used to adjust the sampling weights of inmovers. These two types of household noninterview adjustment factors are represented in the DSE formula as:

f_{cj}^* is the Census Day noninterview adjustment factor
 f_{aj}^* is the A.C.E. interview day noninterview adjustment factor

Each noninterview adjustment spreads the weights of noninterviewed units over interviewed units in the same block cluster and similar type of basic address. There are collapsing rules if the number of interviewed units (in the block cluster by type of basic address category) is too small compared to the number of noninterviewed units.

Interview: A unit is an interview (for the given reference day) if there is at least one data-defined person who possibly or definitely was a resident of the housing unit on the given reference day.

Noninterview: An occupied (as of the given reference date) housing unit that is not an interview is a noninterview.

3.3.2 Characteristic Imputation

P sample characteristic imputation uses either a hot deck method or available demographic distributions (see Cantwell (2000) for details) that imputes race, Hispanic origin, sex, tenure, and age if the item is missing for interviewed households. P sample mover status is not considered when imputing characteristics. Since the E sample is supposed to be a sample from the census, the results of Census 2000 Edit and Imputation are used for persons in the E sample.

3.3.3 Match, Residency and Correct Enumeration Probabilities

Probabilities for persons with unresolved final residence or match status in the P sample or unresolved final correct enumeration status in the E sample are assigned using imputation cell

estimation within groups (see Cantwell 2000 for details). Within each group a probability equal to a simple proportion is imputed for unresolved persons. For example, E sample (or P sample) persons in a group with unresolved enumeration (match) status will be assigned a correct enumeration (match) probability that is the proportion of correct enumerations (matches) among persons with resolved enumeration (match) status in the group. The estimated probabilities are estimated in the DSE formulas as:

Pr_{mj}^* is the estimated match probability for unresolved match status

$Pr_{res,j}^*$ is the estimated residence probability for unresolved residence status

$Pr_{ce,j}^*$ is the estimated correct enumeration probability for unresolved enumeration status

3.4 Targeted Extended Search

Navarro (2000) provides a detailed discussion of the Targeted Extended Search (TES) planned for the Census 2000 A.C.E. survey. From experience with coverage measurement surveys, some A.C.E. block clusters are not going to be counted correctly due to geocoding errors. In both the P and E samples, enumerators will misidentify the location of housing units and include or exclude the units and persons in them incorrectly due to geocoding errors. For the 1990 PES, every block cluster was subject to a complete surrounding block search (one or two rings of surrounding blocks). Expanding the search area (if done in a "balanced" way for the P and E samples) reduced the variance on DSEs. We discovered in the 1990 PES surrounding block search that searching for all unmatched P and E sample persons was not effective; the vast majority of matches in a surrounding block were associated with geocoding errors. For Census 2000, we will perform a TES for one ring of blocks for about 20% of the sample block clusters. Block clusters with an indication from the initial housing unit matching of a large number of geocoding errors will be included in TES with certainty. A sample of the remaining block clusters will be also be included in TES. For DSE weighted correct enumeration rates from the E-sample and match rates from the P-sample will be adjusted to account for the TES sampling. The sampling weights of "TES eligible" persons will include a factor for the TES sampling. In terms of the DSE formula, the inverse of the probability of being selected for TES is incorporated in the sampling weight, W_j . See Navarro (2000) for details.

4. Synthetic Estimation

The estimated coverage correction factors for each post-stratum are used to form synthetic estimates. Synthetic estimation combines coverage error results with census counts at the block level to produce block level population estimates. The synthetic methodology assumes coverage correction factors do not vary within post-stratum by geography. As a result, one coverage correction factor is assumed to be appropriate for all geographic areas within each post-stratum. To obtain block level synthetic estimates, multiply block level census counts for post-strata by

post-strata coverage correction factors. After a controlled rounding technique is implemented, person records are created at the block level. Subsequent tabulations, which are based on the original and created records, are corrected for coverage error. Haines (2000) provides a detailed discussion of synthetic estimation.

REFERENCES

Cantwell, P. (2000), "Accuracy and Coverage Evaluation Survey: Missing Data", DSSD Census 2000 Procedures and Operations Memorandum Series #Q- .

Fay, R. (1988), "Evaluation of Census Coverage Through Direct Measurement of the PEP".

Griffin, R. (1999), "Accuracy and Coverage Evaluation Survey: Post-stratification Research Methodology", DSSD Census 2000 Procedures and Operations Memorandum Series #Q-5.

Griffin, R. and Haines, D. (2000), "Accuracy and Coverage Evaluation Survey: Post-stratification for Dual System Estimation", DSSD Census 2000 Procedures and Operations Memorandum Series #Q- .

Haines, D. (2000), "Accuracy and Coverage Evaluation Survey: Synthetic Estimation", DSSD Census 2000 Procedures and Operations Memorandum Series #Q- .

Hogan, H. (1992), " The 1990 Post-Enumeration Survey: An Overview", The American Statistician, November 1992, Vol. 46, No. 4., 261-269.

Hogan, H. (1993), " The 1990 Post-Enumeration Survey: Operations and Results", Journal of the American Statistical Association, September 1993, Vol. 88, No. 423, 1047-1060.

Killion, R.A. (1998), "Estimation Decisions for the Integrated Coverage Measurement Survey for Census 2000", Census 2000 Decision Memorandum No. 42.

Kostanich, D., Fenstermaker, D., and Griffin, R. (1999), " Accuracy and Coverage Evaluation Survey Plans for Census 2000, Prepared for the March 19, 1999 meeting of the National Academy of Science Panel to Review the 2000 Census.

Navarro, A. (2000), " Accuracy and Coverage Evaluation Survey: Targeted Extended Search Methodology, DSSD Census 2000 Procedures and Operations Memorandum Series #Q- .

Raglin, D. and Bean, S. (1999), " Outmover Tracing and Interviewing", Census 2000 Dress Rehearsal Evaluation Results Memorandum Series #C-3.

Schindler, E. (1999), "Comparison of DSE C and DSE A", Census 2000 Dress Rehearsal Evaluation Memorandum # C-8a.

Wolter, K., "Some Coverage Error Models for Census Data", Journal of the American Statistical Association, June 1986, Vol. 81, No. 394, 338-346.

APPENDIX : THE EFFECT OF MOVERS ON DUAL SYSTEM ESTIMATION

This appendix discusses the effect of movers on Dual system Estimation (DSE). Section 1 describes the alternative methodologies that have been considered by the Census Bureau for dealing with movers for DSE in census coverage measurement surveys. Section 2 presents a derivation and comparison of the correlation bias of DSEs using these alternative methodologies.

1. Alternative Methodologies

There are three alternative methodologies for handling movers in DSE that have been considered by the Census Bureau: These have historically been referred to as PES-C, PES-B, and PES-A however, the current terminology is to refer to them as Procedures A, B, and C. The following are the definitions of these methodologies from Statistical Training document ISP-TR-5, Evaluating Censuses of Population and Housing.

Procedure A: This procedure reconstructs the households as they existed at the time of the census. A respondent is asked to identify all persons who were living or staying in the sample household on census day. These persons are then matched against names on the census questionnaire for the sample address (and surrounding area). From this information, estimates of the number and percent matched for non-movers and out-movers can be made.

Procedure B: This procedure identifies all current residents living or staying in the sample household at the time of the interview. The respondent is asked to provide the address(es) where these persons were living or staying on census day. These persons are then matched against names on corresponding census questionnaire(s) at the non-movers or in-movers census address. Estimates of the number and percent matched for non-movers and in-movers can be made.

Procedure C: This procedure identifies all current residents living or staying at the sample address at the time of the interview plus all other persons who lived at the sample address on census day and have moved since census day. However, only the census day residents (non-movers and out-movers) are matched with the census questionnaire(s) at the sample address. Estimates of the number of non-movers, out-movers, in-movers, and the percent matched for non-movers and out-movers, can then be made. Estimates of non-movers and movers come from Procedure B and match rate estimates for the movers from Procedure A (using out-mover matching). Thus, Procedure C is a combination of Procedure A and B.

In 1990, Procedure B was used. The unresolved match rate for in-movers in 1990 was high, around 13%. In addition with sampling for nonresponse initially planned for Census 2000 in- mover matching would have had an even higher level of difficulty. A decision was made that Procedure B would NOT be used for Census 2000. When the Supreme Court decided against sampling for apportionment it was too late to change the decision on Procedure B.

In the 1995 and 1996 Census tests, Procedure A was used. We felt that a out-mover match rate would be more accurate than a in-mover match rate particularly with sampling for nonresponse. For out-movers, the interviewer attempted a proxy interview to obtain their name and new addresses and data that could be used for matching. Then an attempt could be made to trace the people to obtain an interview with a household member. The best available data for out-movers was matched to their census day address in the same manner as used for the non-movers. Out-mover tracing had problems in 1995 and was tested in 1996 and in the Census 2000 Dress Rehearsal. The outmover tracing evaluation showed that there is little gain in an outmover tracing operation (Raglin (1999)). A decision was made to use the outmover proxy interview data for outmover matching for Census 2000.

We tested Procedure C in the Dress Rehearsal and it will be used in Census 2000 (Schindler (1999)). The advantage of Procedure C is that the estimate of the number of movers uses in-mover data which is more reliable since it is collected from the in-movers themselves. The match rate of the movers is estimated using the out-mover match rate so that the difficulties of in-mover matching are avoided. Out-mover tracing is a problem, however, and in many cases it is necessary to use proxy data for matching (there will be no outmover tracing for Census 2000). Procedure C is an attempt to obtain a Procedure B estimate with no in-mover matching. Procedure C and Procedure B estimates are different since out-movers do not have the same match rate as in-movers. However, the Procedure B in-mover match rate estimate is unreliable due to a high percentage of unresolved cases.

2. Correlation Bias

For a given post-stratum there are N persons in the true population. As in Wolter (1986), we have the following 2x2 table of inclusion probabilities for person i :

	In Census	Out of Census	Total
In A.C.E.	P_{i11}	P_{i12}	P_{i1+}
Out of A.C.E.	P_{i21}	P_{i22}	P_{i2+}
Total	P_{i+1}	P_{i+2}	P_{i++}

$$DSE = N_{++} = \frac{(N_{+1})(N_{1+})}{N_{11}}$$

The leading term in the bias of the DSE from Wolter (1986) is given as equation (1) for N persons in a post-stratum where p_{i1+} is the A.C.E. capture probability and p_{i+1} is the Census capture probability.:

$$\overline{p_{1+}} = N^{-1} \sum_i p_{i1+}$$

$$\text{Bias (DSE)} = - \frac{N\sigma(p_{1+}, p_{+1})}{\sigma(p_{1+}, p_{+1}) + \overline{p_{1+}} \overline{p_{+1}}} \quad (1)$$

$$\overline{p_{+1}} = N^{-1} \sum_i p_{i+1}$$

$$\overline{p_{1+}p_{+1}} = N^{-1} \sum_i p_{i1+}p_{i+1}$$

$$\text{where, } \sigma(p_{1+}, p_{+1}) = \overline{p_{1+}p_{+1}} - \overline{p_{1+}} \overline{p_{+1}}$$

Assume for a post-stratum there are only two different capture probabilities for the Census and for A.C.E; one for movers (M) and one for nonmovers (N). Furthermore

$$\begin{aligned} p_{+1M} &= cp_{+1N} \text{ (Census)} \\ p_{1+M} &= mp_{1+N} \text{ (A.C.E.)} \\ N &= N_M + N_N \\ N_M &= dN_N \\ 0 &< c < 1 \\ 0 &< m < 1 \end{aligned}$$

Then it can be shown that:

$$\overline{p_{1+}p_{+1}} = \frac{(p_{1+N}p_{+1N})(1+dc m)}{d+1}$$

$$\overline{p_{1+}} = \frac{p_{1+N}(1+dm)}{d+1}, \text{ so (1) is as follows}$$

Note that if $c=1$ or $m=1$ we have homogeneity in one system and Bias (DSE) = 0.

$$\overline{p_{+1}} = \frac{p_{+1N}(1+dc)}{d+1}$$

$$\text{Bias(DSE)} = -\frac{Nd(1-c)(1-m)}{(1+dc)(d+1)}$$

Now for a given c and d , the absolute value of Bias (DSE) increases as m decreases.

Now assume for inmovers (IM) $m = b$ and for outmovers (OM) $m = a$. $a < b$ is reasonable as we know we have more difficulty capturing outmovers than inmovers. Thus the absolute bias of DSE is greater for outmovers (Procedure A) than inmovers (Procedure B).

For the following table N_{ij} is the total number of the N persons falling in cell ij .

Using a N subscript for nonmovers, a O subscript for outmovers, and a I subscript for inmovers, the DSE for Procedure C is as follows:

	In Census	Out of Census	Total
In A.C.E.	N_{11}	N_{12}	N_{1+}
Out of A.C.E.	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N_{++}

$$DSE_C = \frac{N_{+1}(N_{1+N} + N_{1+I})}{N_{11N} + N_{1+I} \left(\frac{N_{11O}}{N_{1+O}} \right)}$$

Assume that,

$$E\left(\frac{N_{11O}}{N_{1+O}}\right) = \frac{N_I \overline{P_{1+I} P_{+1I}}}{N_I \overline{P_{1+I}}}$$

then the leading term of the expected value of DSE_C is given by:

$$E(DSE_C) = \frac{\overline{NP_{+1}} (N_N \overline{P_{1+N}} + N_I \overline{P_{1+I}})}{N_N \overline{P_{1+N} P_{+1N}} + N_I \overline{P_{1+I} P_{+1I}}} = E(DSE_B)$$

Thus DSE_C has about the same bias as DSE_B if the outmover match rate is a good estimate of the inmover match rate.

cc: DSSD Census 2000 Procedures and Operations Memorandum Series Distribution List



MASTER FILE

January 12, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-22

MEMORANDUM FOR

Howard Hogan
Chief, Decennial Statistical Studies Division

From:

Donna Kostanich *DK*
Assistant Division Chief, Sampling and Estimation
Decennial Statistical Studies Division

Prepared by:

Dawn Haines *del*
Estimation Team

Subject:

Accuracy and Coverage Evaluation Survey: Synthetic Estimation

I. INTRODUCTION

This memorandum documents the Accuracy and Coverage Evaluation (A.C.E.) methodology associated with synthetic estimation. The Census Bureau uses synthetic estimation to provide population estimates for small geographic areas such as blocks, tracts, counties, and congressional districts. Synthetic estimates are formed by combining coverage measurement results with census counts to produce population estimates for any geographic area of interest. For example, a block-level synthetic estimate is formed by distributing a post-stratum's coverage correction factor to blocks proportional to the size of the post-stratum's population within the block. Rounded, adjusted synthetic estimates at the tabulation block level constitute the PL 94-171 redistricting file which is used to produce all census data products.

The synthetic methodology assumes that coverage correction factors are uniform within a given post-stratum. Another way of saying this is that the coverage error rate for a given post-stratum is the same within all blocks. To the extent that the synthetic assumption fails, the estimates of coverage for individual areas will be biased. Failure of the synthetic assumption leads to synthetic estimation bias in population size estimates for a given area. Synthetic estimation bias decreases as the size of geography increases.

Essentially, the synthetic estimation methodology for the Census 2000 A.C.E. is the same as that used for the 1990 Post Enumeration Survey (PES). There will be more levels of controlled rounding in the Census 2000 A.C.E. than there was in the 1990 PES. Barring computer resource limitations, the proposed geographic levels of controlled rounding for the Census 2000 A.C.E. are post-stratum, state, county, tract, and block. The increased number of geographic levels represents an improvement over the 1990 methodology since rounded synthetic estimates will be

similar to the unrounded synthetic estimates for all levels of controlled rounding. Another methodological change for synthetic estimation relates to the coverage correction factors. Coverage correction factors for the 1990 census were based on tabulation block geography and applied synthetically to 1990 tabulation block counts. In contrast, Census 2000 coverage correction factors will be based on collection block geography and applied synthetically to 2000 tabulation block counts.

This memorandum describes the calculation of synthetic estimates. Also, the controlled rounding procedure used to produce integer-valued synthetic estimates is described in 11 steps. Attachment 1 provides a corresponding visual representation of the 11 steps in the controlled rounding process.

II. SYNTHETIC ESTIMATION

A. Calculation

Consider forming synthetic estimates for geographic level g for a given post-stratum. Let $C_{i,g}$ denote the census count for post-stratum i in geographic level g and define CCF_i to be the coverage correction factor for post-stratum i . The general form for a synthetic estimate for post-stratum i at geographic level g is calculated as

$$\hat{N}_{i,g}^s = C_{i,g} \times CCF_i.$$

Aggregating synthetic estimates over all the post-strata in geographic level g yields a synthetic estimate for geographic level g . This is denoted as

$$\hat{N}_g^s = \sum_i C_{i,g} \times CCF_i.$$

The eventual goal of synthetic estimation and the controlled rounding procedure is to produce rounded, adjusted synthetic estimates at the tabulation block level. These estimates comprise the PL 94-171 redistricting file which is the source file for all census data products.

B. Geography

It is important to note that the components of a synthetic estimate are not necessarily based on identical geography. Specifically, synthetic estimation census counts are based on **tabulation block** geography while the coverage correction factors are based on **collection block** geography. Although this appears to be a minor detail, it could have important ramifications on variables with a geographic component.

For example, consider the post-stratification variable mail return rate. Mail return rate is calculated at the tract level and is based on collection tract definitions. People are assigned to post-strata based on the mail return rate of collection tracts. Now consider the case where people are assigned to post-strata based on the mail return rate of tabulation tracts. It could be the case that the change in geography causes an individual's post-stratum assignment to change. For example, suppose the mail return rate of a collection tract is 80 percent. Also, suppose the collection tract is split into two pieces by a tabulation tract. A person once belonging to a collection tract with an 80 percent mail return rate may now belong to a tabulation tract with a different mail return rate. Changes in an individual's post-stratum causes the dual system estimates, coverage correction factors, and synthetic estimates to also change.

To avoid potential inconsistencies in the assignment of people to post-strata, there will be only one assignment of people to post-strata. The assignment will be based on collection block geography. Further, this assignment will be maintained for all estimation purposes.

C. Adjustment

Synthetic estimates at any geographic level are not typically integer-valued. A controlled rounding program, developed by the Statistical Research Division (SRD) of the U.S. Bureau of the Census, is utilized which produces integer-valued estimates. In essence, the controlled rounding program takes a two-dimensional matrix of numbers and rounds each to an adjacent integer value based on an efficiency algorithm. The two dimensions of the matrix are the post-strata for one level of geography by totals for a lower level of geography. The controlled rounding procedure ensures that the sum of the synthetic estimates within a geographic level are not rounded up or down by more than one.

The overall goal of controlled rounding is to obtain an integer number of persons for each post-strata i within each tabulation block b , representing overcounts and undercounts. The controlled rounding program cannot be implemented in one step due to the size of the post-strata by tabulation block matrix. As a result, controlled rounding is implemented in steps such that the rounded, adjusted synthetic estimates at block level b :

- (1) sum to the rounded, adjusted synthetic estimates at tract level t , and
- (2) sum to the rounded, adjusted synthetic estimates at county level c , and
- (3) sum to the rounded synthetic estimates at state level s .

The controlled rounding procedure, which is diagramed in Attachment 1, is implemented as follows:

1. Form synthetic estimates for Post-stratum i within State s , $\hat{N}_{i,s}^S$. The letter S denotes a synthetic estimate.

2. Apply the controlled rounding procedure to state-level synthetic estimates to produce rounded, state-level synthetic estimates, denoted $\hat{N}_{i,s}^{RS}$. The letters *RS* denote a rounded, synthetic estimate. The two dimensions of this matrix are State *s* by Post-stratum *i*.
3. Form the ratio of the rounded state-level synthetic estimate to the state-level synthetic estimate for Post-stratum *i* in State *s*.
4. For each Post-stratum *i* within County *c* for State *s*, multiply the county-level synthetic estimate by the ratio formed in step 3. The letters *AS* denote an adjusted, synthetic estimate. The resulting product is the adjusted county-level synthetic estimate for Post-stratum *i*, written as

$$\hat{N}_{i,c}^{AS} = \hat{N}_{i,c}^S \left[\frac{\hat{N}_{i,s}^{RS}}{\hat{N}_{i,s}^S} \right] \text{ where } \hat{N}_{i,c}^S = C_{i,c} \times CCF_i.$$

5. Apply the controlled rounding procedure to the adjusted county-level synthetic estimates to produce rounded, adjusted, county-level synthetic estimates, denoted $\hat{N}_{i,c}^{RS}$. The two dimensions of this matrix are County *c* in State *s* by Post-stratum *i* in State *s*.
6. Form the ratio of the rounded, adjusted, county-level synthetic estimate to the county-level synthetic estimate for Post-stratum *i* in County *c* in State *s*.
7. For each Post-stratum *i* within Tract *t* in County *c* for State *s*, multiply the tract-level synthetic estimate by the ratio formed in step 6. The resulting product is the adjusted tract-level synthetic estimate for Post-stratum *i*, written as

$$\hat{N}_{i,t}^{AS} = \hat{N}_{i,t}^S \left[\frac{\hat{N}_{i,c}^{RS}}{\hat{N}_{i,c}^S} \right] \text{ where } \hat{N}_{i,t}^S = C_{i,t} \times CCF_i.$$

8. Apply the controlled rounding procedure to the adjusted tract-level synthetic estimates to produce rounded, adjusted tract-level synthetic estimates, denoted $\hat{N}_{i,t}^{RS}$. The two dimensions of this matrix are Tract *t* in County *c* in State *s* by Post-stratum *i* in County *c* in State *s*.
9. Form the ratio of the rounded, adjusted tract-level synthetic estimate to the tract-level synthetic estimate for Post-stratum *i* in Tract *t* in County *c* in State *s*.

10. For each Post-stratum i within Block b in Tract t in County c for State s , multiply the block-level synthetic estimate by the ratio formed in step 9. The resulting product is the adjusted block-level synthetic estimate for Post-stratum i , written as

$$\hat{N}_{i,b}^{AS} = \hat{N}_{i,b}^S \left[\frac{\hat{N}_{i,t}^{RS}}{\hat{N}_{i,t}^S} \right] \quad \text{where} \quad \hat{N}_{i,b}^S = C_{i,b} \times CCF_i.$$

11. Again, apply the controlled rounding procedure to the adjusted block-level synthetic estimates to produce rounded, adjusted block-level synthetic estimates, denoted $\hat{N}_{i,b}^{RS}$. The two dimensions of this matrix are Block b in Tract t in County c in State s by Post-stratum i in Tract t in County c in State s .

III. RECORD CREATION FOR COVERAGE CORRECTION

Once the rounded, adjusted block-level synthetic estimates are formed, they are compared with the census counts for post-stratum i in tabulation block b . Person records are then created at the post-stratum level to reflect the coverage correction for the census blocks. The number of records to create depends on the value of the coverage correction factor.

A. Coverage Correction Factors ≥ 1

Define the integer

$$U_{i,b} = \hat{N}_{i,b}^{RS} - C_{i,b}$$

for each post-stratum i in tabulation block b . If $U_{i,b} = 0$, then no additional records are created. If $U_{i,b} > 0$, then create $U_{i,b}$ undercount person records for post-stratum i in tabulation block b .

Additional person records are created by randomly selecting without replacement $U_{i,b}$ records from the $C_{i,b}$ available person records in post-stratum i and tabulation block b . The selected records are replicated and appended to the file of person records. All housing unit and person identifier information is maintained for each record. A value will be assigned to the GQ Type field to signify an undercount person record for each of the replicated records. This results in an upward adjustment of people in post-stratum i in tabulation block b .

B. Coverage Correction Factors < 1

Define the integer

$$O_{i,b} = C_{i,b} - \hat{N}_{i,b}^{RS}$$

for each post-stratum i in tabulation block b . If $O_{i,b} > 0$, then create $O_{i,b}$ overcount person records for post-stratum i in tabulation block b .

Overcount person records are created by randomly selecting without replacement $O_{i,b}$ records from the $C_{i,b}$ available person records in post-stratum i and tabulation block b . The selected records are replicated and appended to the file of person records. All housing unit and person identifier information is maintained for each record. A value will be assigned to the GQ Type field to signify an overcount person record for each of the replicated records. This results in a downward adjustment of people in post-stratum i in tabulation block b .

cc: DSSD Census 2000 Procedures and Operations Distribution List
Statistical Design Program Steering Committee Team Leaders
A.C.E. Sample Design Team
A.C.E. Estimation Team

Attachment 1: Stages of Controlled Rounding

1. Form synthetic estimates for Post-stratum i within State s:

$$\hat{N}_{i,s}^S = C_{i,s} \times CCF_i$$

2. Apply controlled rounding procedure to $\hat{N}_{i,s}^S$:

Post-stratum i					
State	1	2	..	i	..
1	$\hat{N}_{i,s}^S$				
2					
:					
s					
:					

→

Post-stratum i					
State	1	2	..	i	..
1	$\hat{N}_{i,s}^{RS}$				
2					
:					
s					
:					

3. & 4. Form adjusted synthetic estimate for Post-stratum i within County c for State s:

$$\hat{N}_{i,c}^{AS} = \hat{N}_{i,c}^S \left[\frac{\hat{N}_{i,s}^{RS}}{\hat{N}_{i,s}^S} \right] \quad \text{where} \quad \hat{N}_{i,c}^S = C_{i,c} \times CCF_i$$

5. Apply controlled rounding procedure to $\hat{N}_{i,c}^{AS}$:

Post-stratum i in State s					
County	1	2	..	i	..
1	$\hat{N}_{i,c}^{AS}$				
2					
:					
c					
:					

→

Post-stratum i in State s					
County	1	2	..	i	..
1	$\hat{N}_{i,c}^{RS}$				
2					
:					
c					
:					

6. & 7. Form adjusted synthetic estimate for Post-stratum i within Tract t in County c for State s:

$$\hat{N}_{i,t}^{AS} = \hat{N}_{i,t}^S \left[\frac{\hat{N}_{i,c}^{RS}}{\hat{N}_{i,c}^S} \right] \quad \text{where} \quad \hat{N}_{i,t}^S = C_{i,t} \times CCF_i$$

8. Apply controlled rounding procedure to $\hat{N}_{i,t}^{AS}$:

Post-stratum i in County c in State s					
Tract	1	2	..	i	..
1	$\hat{N}_{i,t}^{AS}$				
2					
:					
t					
:					

→

Post-stratum i in County c in State s					
Tract	1	2	..	i	..
1	$\hat{N}_{i,t}^{RS}$				
2					
:					
t					
:					

9. & 10. Form adjusted synthetic estimate for Post-stratum i within Block b in Tract t in County c for State s:

$$\hat{N}_{i,b}^{AS} = \hat{N}_{i,b}^S \left[\frac{\hat{N}_{i,t}^{RS}}{\hat{N}_{i,t}^S} \right] \quad \text{where} \quad \hat{N}_{i,b}^S = C_{i,b} \times CCF_i$$

11. Apply controlled rounding procedure to $\hat{N}_{i,b}^{AS}$:

Post-stratum i in Tract t in County c in State s					
Block	1	2	..	i	..
1	$\hat{N}_{i,b}^{AS}$				
2					
:					
b					
:					

→

Post-stratum i in Tract t in County c in State s					
Block	1	2	..	i	..
1	$\hat{N}_{i,b}^{RS}$				
2					
:					
b					
:					



MASTER FILE

April 19, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM SERIES #Q-24

MEMORANDUM FOR Howard Hogan
 Chief, Decennial Statistical Studies Division

From: Donna Kostanich *DK*
 Assistant Division Chief, Sampling and Estimation
 Decennial Statistical Studies Division

Prepared by: Richard Griffin *RG*
 Dawn Haines *leh*
 Estimation Staff

Subject: Accuracy and Coverage Evaluation Survey: Final Post-
 stratification Plan for Dual System Estimation

I. INTRODUCTION

The goal of post-stratification is to group together people who have similar coverage by the census. A common assumption is that people who share similar housing, similar language, similar cultural attitudes, and similar education would also share similar census coverage. Tenure, race and ethnic origin often serve as a marker for these similarities.

This memorandum presents the final post-stratification plan for the Accuracy and Coverage Evaluation (A.C.E.) Survey including Puerto Rico. The plan for Census 2000 A.C.E. is summarized in Section III. The detailed definitions of the poststratification variables and the race and Hispanic origin groups are given in Sections IV. and V., respectively.

II. BACKGROUND

The 2000 A.C.E. is different from the 1990 Post Enumeration Survey (PES). The A.C.E. will have approximately twice the sample size of the PES. This larger sample size allows for the formation of more post-strata and more post-strata have the advantage of reducing correlation bias. Additionally in 2000 multiple responses to the race question will be permitted; whereas in 1990 only one race could be selected.

The 1990 PES had 357 post-strata defined by a cross-classification of 51 post-stratum groups by seven age/sex groups. The 357 design started with a cross-classification of seven variables: age,

sex, race, Hispanic origin, tenure, urbanicity, and region. There were 840 cells in the cross-classification. Collapsing was necessary in order to produce post-strata with sufficient sample for reliable Dual System Estimation (DSE). The attachment shows the 51 post-stratum groups for the 1990 PES after collapsing and the seven age/sex groups.

Race and Hispanic origin were the most important variables. After collapsing, five race/Hispanic origin post-strata were maintained: Non-Hispanic White or Other, Black, Hispanic White or Other, Asian and Pacific Islander, and Reservation Indians. Off-reservation American Indians were placed in either the Non-Hispanic White or Other group or the Hispanic White or Other group depending on whether they were of Hispanic origin. Within each of these race/Hispanic origin post-strata, seven age/sex categories were maintained.

The other variables were collapsed in the following order: region, urbanicity, then tenure, if necessary. For American Indians residing on reservations, all these variables were collapsed. For Asian and Pacific Islanders, region and urbanicity were collapsed and tenure maintained. For the Black and Hispanic White or Other groups, region was collapsed for two levels of urbanicity. For Non-Hispanic White or Other, the full cross-classification of region, urbanicity and tenure were maintained.

The 1990 PES for Puerto Rico had 21 post-strata defined by a cross-classification of 3 Place Type categories and seven age/sex categories. The place types were central city areas in Metropolitan Statistical Areas, non-central cities in Metropolitan Statistical Areas, and areas outside of Metropolitan Statistical Areas. The seven age/sex categories were the same as those used for the U.S. These 1990 post-stratification groups for Puerto Rico are also given in the attachment.

III. CENSUS 2000 A.C.E. POST-STRATIFICATION PLAN

For the Census 2000 A.C.E. we will retain most of the 1990 PES post-stratification variables and we will include several additional variables. The 2000 A.C.E. post-strata will be defined by nine variables: age, sex, race, Hispanic origin, tenure, region, Metropolitan Statistical Area size, Type of Enumeration Area, and tract level return rate. The Metropolitan Statistical Area size variable is replacing the urbanicity variable which will not be available until the summer of 2001. Type of Enumeration Area and the tract return rate are two new features of the 2000 A.C.E. post-stratification. The mailout/mailback areas will be differentiated from other types of enumeration areas. Tracts will be classified by high or low return rate. Additionally, multiple responses to the race question will be reflected in the race and Hispanic origin groupings.

Table 1a shows the 64 post-stratum groups for the Census 2000 A.C.E.. Within each post-stratum group there will be seven age/sex groups (Table 1c). Thus, there is a maximum of $64 \times 7 = 448$ post-strata, and of course there will be fewer if further collapsing is necessary. The post-stratification plan was chosen to reduce correlation bias without having an adverse effect on the variance of the Dual System Estimator.

For the Census 2000 A.C.E. in Puerto Rico, post-strata will be defined by five variables: age, sex, tenure, Metropolitan Statistical Area, and tract-level return rate. The variable region is not applicable for Puerto Rico. Further, there is only one Type of Enumeration Area (Update/Leave) in Puerto Rico, so this variable is not utilized. Table 1b shows the 12 post-stratum groups used in the Puerto Rico Census 2000 A.C.E. Survey. Within each post-stratum group, the seven age/sex groups in Table 1c are utilized. Thus, there is a maximum of $12 \times 7 = 84$ post-strata, and of course there will be fewer if further collapsing is necessary.

Tables 1a and 1b show the 64 and 12 post-stratum groups for the U.S. and Puerto Rico, respectively. Table 1c presents the seven age/sex groups which are used for both the U.S. and Puerto Rico. Subsequent sections of this memorandum provide a detailed description of the post-stratification domains and variables, including any alternative definitions for Puerto Rico. An extensive explanation of the domains is presented in Section V.

Table 1a: Census 2000 A.C.E. - 64 Post-Stratum Groups (U.S.)

Race/Hispanic Origin Domain Number*		Tenure	MSA/TEA	High Return Rate				Low Return Rate			
				N	M	S	W	N	M	S	W
Domain 7 (Non-Hispanic White or "Some other race")		Owner	Large MSA MO/MB	1	2	3	4	5	6	7	8
			Medium MSA MO/MB	9	10	11	12	13	14	15	16
			Small MSA & Non-MSA MO/MB	17	18	19	20	21	22	23	24
			All Other TEAs	25	26	27	28	29	30	31	32
		Non-owner	Large MSA MO/MB	33				34			
			Medium MSA MO/MB	35				36			
			Small MSA & Non-MSA MO/MB	37				38			
			All Other TEAs	39				40			
Domain 4 (Non-Hispanic Black)		Owner	Large MSA MO/MB	41				42			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	43				44			
			All Other TEAs								
		Non-owner	Large MSA MO/MB	45				46			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	47				48			
			All Other TEAs								
Domain 3 (Hispanic)		Owner	Large MSA MO/MB	49				50			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	51				52			
			All Other TEAs								
		Non-owner	Large MSA MO/MB	53				54			
			Medium MSA MO/MB								
			Small MSA & Non-MSA MO/MB	55				56			
			All Other TEAs								
Domain 5 (Native Hawaiian or Pacific Islander)		Owner	57								
		Non-owner	58								
Domain 6 (Non-Hispanic Asian)		Owner	59								
		Non-owner	60								
American Indian or Alaska Native	Domain 1 (On Reservation)	Owner	61								
		Non-owner	62								
	Domain 2 (Off Reservation)	Owner	63								
		Non-owner	64								

* For Census 2000 persons can self identify with more than one race group. For post-stratification, persons are included in a single Race/Hispanic Origin domain. This does not change a person's actual response and all persons will be tabulated based on their actual response in the census. An extensive explanation of the domains is presented in Section V.

Table 1b: Census 2000 A.C.E. - 12 Post-Stratum Groups (Puerto Rico)

Tenure	MSA	High Return Rate	Low Return Rate
Owner	San Juan CMSA	1	2
	Other MSA	3	4
	Non-MSA	5	6
Non-owner	San Juan CMSA	7	8
	Other MSA	9	10
	Non-MSA	11	12

Table 1c: Census 2000 A.C.E. - 7 Age/Sex Groups (U.S. & Puerto Rico)

	Male	Female
Under 18	A	
18 to 29	B	C
30 to 49	D	E
50+	F	G

Key:

Return Rate: Tract-level variable measuring the proportion of occupied housing units in the mailback universe which returned a census questionnaire. Low return rate tracts are those tracts whose return rate is less than or equal to the 25th percentile return rate.

MSA: Metropolitan Statistical Area or Consolidated Metropolitan Statistical Area, as defined by the Office of Management and Budget (OMB), will be referred to as MSA.

TEA: Type of Enumeration Area.

MO/MB: Mailout/Mailback Type of Enumeration Area.

N, M, S, W: Refers to region - Northeast, Midwest, South, West.

"Some other race": One of six possible major race categories obtained from the census questionnaire.

Further details on the variable definitions are included in the following sections.

IV. CENSUS 2000 POST-STRATIFICATION VARIABLES

A. Post-stratification Variables

A.C.E. post-stratification will use the following variables:

- Race/Hispanic Origin - seven categories (omitted for Puerto Rico)
- Age/Sex - seven categories
- Tenure - two categories
- Metropolitan Statistical Area (MSA) by Type of Enumeration Area (TEA) - four categories (three categories for Puerto Rico)
- Return Rate - two categories
- Region - four categories (omitted for Puerto Rico)

The seven Race/Hispanic Origin domains are:

- American Indian or Alaska Native on Reservations
- Off-Reservation American Indian or Alaska Native
- Hispanic
- Non-Hispanic Black
- Native Hawaiian or Pacific Islander
- Non-Hispanic Asian
- Non-Hispanic White or "Some other race"

See Section V. for further details on the Race/Hispanic Origin domains. Inclusion in a Race/Hispanic Origin domain is complicated as it depends on several variables and whether there are multiple race responses. In addition, inclusion in a Race/Hispanic Origin domain **does not** change a persons Race/Hispanic Origin response. All Census 2000 tabulations will be based on the actual responses. For example, a person who responds as American Indian on a reservation and Black will be placed in the first Race/Hispanic Origin domain (Group 1) for post-stratification purposes but will be tabulated in the census as American Indian/Black.

The seven Age/Sex categories are:

- Under 18
- 18 - 29 Male
- 18 - 29 Female
- 30 - 49 Male
- 30 - 49 Female
- 50+ Male
- 50+ Female

The two Tenure categories are:

- Owner
- Non-owner

The four MSA/TEA categories are:

- Large MSA Mailout/ Mailback (MO/MB)
- Medium MSA MO/MB
- Small MSA or Non-MSA MO/MB
- All other TEAs

MSA/CMSA FIPS codes, as defined by the Office of Management and Budget (OMB), will be used for post-stratification. For simplification, MSA/CMSA will herein be referred to as MSA. Large MSA consists of the ten largest MSAs based on unadjusted, Census 2000 total population counts including the population in Group Quarters. Medium MSAs are those (besides the largest 10) which have at least 500,000 total population. Small MSAs are those with a total population size strictly less than 500,000. For post-stratification purposes, MO/MB areas are contrasted with the non-MO/MB areas.

For Puerto Rico there are three MSA categories. The TEA portion of this variable is nonexistent since all of Puerto Rico is Update/Leave. The three MSA categories are:

- San Juan CMSA (San Juan-Caguas-Arecibo CMSA)
- Other MSA (Aguadilla, Mayaguez, and Ponce MSAs)
- Non-MSA

The two Return Rate categories are:

- High
- Low

Return rate is a tract-level variable measuring the proportion of occupied housing units in the mailback universe which returned a census questionnaire. Low (high) return rate tracts are those tracts whose return rate is less than or equal to (greater than) the 25th percentile return rate. Separate 25th percentile cut-off values will be formed for the six applicable Race/Hispanic Origin by Tenure groups. Persons in List/Enumerate, Rural Update/Enumerate, and Urban Update/Enumerate TEAs are automatically placed in the High category. For Puerto Rico, distinct 25th percentile return rate cut-off values will be formed for each Tenure category.

The four Region categories are:

- Northeast
- Midwest
- South
- West

B. Pre-collapsing

All Race/Hispanic Origin, Age/Sex, and Tenure categories for the U.S. will initially be maintained. The pre-collapsing plan for Region, MSA/TEA and Return Rate varies as follows:

- Non-Hispanic White or "Some other race" Owners: No collapsing
- Non-Hispanic White or "Some other race" Non-owners: Eliminate Region
- Non-Hispanic Black: Eliminate Region and partial collapsing of the MSA/TEA variable within Return Rate and Tenure categories
- Hispanic: Eliminate Region and partial collapsing of the MSA/TEA variable within Return Rate and Tenure categories
- Native Hawaiian or Pacific Islander: Eliminate the Region, Return Rate and MSA/TEA variables (Retain Tenure and Age/Sex only)
- Non-Hispanic Asian: Eliminate the Region, Return Rate and MSA/TEA variables (Retain Tenure and Age/Sex only)
- American Indian or Alaska Native on Reservations: Eliminate the Region, Return Rate and MSA/TEA variables (Retain Tenure and Age/Sex only)
- Off-Reservation American Indian or Alaska Native: Eliminate the Region, Return Rate and MSA/TEA variables (Retain Tenure and Age/Sex only)

For Puerto Rico, all 84 post-strata defined by MSA, Tenure, Return Rate, and Age/Sex will initially be maintained. Thus, there will be no pre-collapsing for Puerto Rico.

C. Post-collapsing

Depending on the actual A.C.E. sample sizes, additional collapsing may be necessary. The collapsing procedure is hierarchical which requires a pre-defined collapsing order. Given the pre-collapsing plan which yielded 448 post-strata, not much post-collapsing is anticipated. However, an extensive post-collapsing strategy is presented for completeness and to satisfy the requirement of pre-specification.

Note that collapsing does not necessarily imply elimination of a variable. Collapsing can refer to a reduction in the number of categories for a variable. For both the U.S. and Puerto Rico, a post-stratum is deemed too small if it contains fewer than 100 P Sample persons. The following general outline describes the post-collapsing hierarchy which is applied to both the U.S. and Puerto Rico. Any differences in definitions for Puerto Rico are noted.

If any of the 448 U.S. or 84 Puerto Rico post-strata are too small, Age/Sex will be collapsed first. This means that within any of the 64 U.S. (or 12 Puerto Rico) post-stratum groups, the seven

Age/Sex categories defined in Table 1c will be reduced to the following three categories: Under 18, 18+ Male, and 18+ Female.

If some post-strata are still too small and require collapsing, Region will be collapsed next, if applicable. This collapsing applies only to the Non-Hispanic White or "Some other race" domain since the variable Region is only included in their post-stratification definition. In this case, all levels of Region (Northeast, Midwest, South, West) will be combined to eliminate the variable.

Next, the four-level MSA/TEA variable in the U.S. will be collapsed, if necessary, into the following two groups:

- Large and Medium MSA MO/MB
- Small MSA and Non-MSA MO/MB and All Other TEAs

For Puerto Rico, the three-level MSA variable will be collapsed, if necessary, into the following two groups:

- San Juan CMSA
- Other MSA and Non-MSA

If further collapsing is necessary, Return Rate is the next variable to collapse. High and Low Return Rate categories are combined to eliminate the variable.

Further collapsing involves the variable MSA/TEA in the U.S. (MSA in Puerto Rico). If necessary, the two groups defined above would be combined together to eliminate the variable MSA/TEA for the U.S. (MSA in Puerto Rico) completely.

The next variable to collapse is Tenure. Owner and Non-owner categories are combined to eliminate the variable entirely, if necessary.

If collapsing is still needed, the three remaining Age/Sex post-strata will be combined together to eliminate the Age/Sex variable completely.

In the event that there are not at least 100 P Sample persons in a Race/Hispanic Origin domain, all persons in that domain will be combined with Domain 7, which includes Non-Hispanic White and "Some other race."

V. RACE AND HISPANIC ORIGIN CLASSIFICATIONS

The Census 2000 questionnaire has 15 possible race responses. The 15 responses are collapsed into six major race groups as shown below. Races which are included in the major groups are shown in parentheses. Persons self-identifying with a single race essentially place themselves into one of these six categories.

- White
- Black (Black, African American, Negro)
- American Indian or Alaska Native
- Asian (Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian)
- Native Hawaiian or Pacific Islander (Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander)
- "Some other race"

For the first time in census history, persons will be able to respond to more than one race category. Allowing persons to self-identify with multiple races results in many more than six race groups. In fact, after collapsing race to the six major groups, there are $2^6 - 1 = 63$ possible race combinations. It is necessary to subtract the 1 in this equation since each individual is assumed to have a race.

The race variable defined above is often cross-classified with the Hispanic origin variable to define post-strata. The Hispanic origin variable consists of two responses, No and Yes. Categories which are included in the Yes response are shown in parentheses.

- No, not Spanish/Hispanic/Latino
- Yes (Mexican, Mexican American, Chicano, Puerto Rican, Cuban, Other Spanish/Hispanic/Latino)

Combining the race and Hispanic origin variables yields $63 \times 2 = 126$ possible Race/Hispanic Origin groups. It is important to note that any post-stratification plan of interest cannot support 126 Race/Hispanic Origin groups. As a solution, each of the 126 Race/Hispanic Origin response possibilities are assigned to one of seven Race/Hispanic Origin domains. The seven Race/Hispanic Origin domains are defined as follows:

- American Indian or Alaska Native on Reservations
- Off-Reservation American Indian or Alaska Native
- Hispanic
- Non-Hispanic Black
- Native Hawaiian or Pacific Islander
- Non-Hispanic Asian
- Non-Hispanic White or "Some other race"

Note that missing race and Hispanic origin data are imputed. Rules for classifying the 126 race and Hispanic origin combinations into one of the seven Race/Hispanic Origin domains are now presented. Many of the decisions on how to classify multiple race persons are based on cultural,

linguistic, and sociological factors which are known to affect coverage and are not necessarily data-driven.

A hierarchy is used to assign persons to a Race/Hispanic Origin domain. The Race/Hispanic Origin designation occurs in the following order: American Indian or Alaska Native on Reservations, Off-Reservation American Indian or Alaska Native, Hispanic, Non-Hispanic Black, Native Hawaiian or Pacific Islander, Non-Hispanic Asian, and Non-Hispanic White or "Some other race." All census data are tabulated using the race and Hispanic origin categories selected by census respondents.

For the following tables, Indian Country (IC) is a block-level variable that indicates whether a collection block is (wholly/partially) inside an American Indian reservation/trust land, Tribal Jurisdiction Statistical Area (TJSA), Tribal Designated Statistical Area (TDSA), or Alaska Native Village Statistical Area (ANVSA).

Tables 2 and 3 display the assignment of Race/Hispanic Origin domains. Table 2 applies to Hispanic persons while Table 3 applies to non-Hispanic persons. The first six rows of Tables 2 and 3 correspond to a single race response. The remaining portion of the tables addresses the assignment of multiple race responses to a single Race/Hispanic Origin domain. Although a person may be associated with multiple race responses, each person is included in only one of the seven Race/Hispanic Origin domains. All persons with a common number are assigned to the same Race/Hispanic Origin domain. Following is a verbal description of who is included in each Race/Hispanic Origin domain and their associated domain number.

Domain 1 (Includes American Indian or Alaska Native on Reservations): This domain includes any person living on a reservation marking American Indian or Alaska Native either as their single race or as one of many races, regardless of their Hispanic origin.

Domain 2 (Includes Off-Reservation American Indian or Alaska Native): This domain includes any person living in IC but not on a reservation who marks American Indian or Alaska Native either as their single race or as one of many races, regardless of their Hispanic origin. This domain also includes any non-Hispanic person not living in IC who marks American Indian or Alaska Native as their single race.

Domain 3 (Includes Hispanic): This domain includes all Hispanic persons who are not included in Domains 1 or 2. All Hispanic persons who self-identify with three or more races (excluding American Indian or Alaska Native in IC) are included in Domain 3. The only exception to this rule occurs when a Hispanic person lives in the state of Hawaii and classifies themselves as Native Hawaiian or Pacific Islander, regardless of whether they identify with a single or multiple race. All Hispanic persons satisfying this condition are re-classified into Domain 5.

Domain 4 (Includes Non-Hispanic Black): This domain includes any non-Hispanic person who marks Black as their only race. It also includes the combination of Black and American Indian or Alaska Native not in IC. In addition, people who mark Black and another single race group (Native Hawaiian or Pacific Islander, Asian, White, or "Some other race") are included in Domain 4. The only exception to this rule occurs when a Non-Hispanic Black person lives in the state of

Hawaii and classifies themselves as Native Hawaiian or Pacific Islander. All Non-Hispanic Black persons satisfying this condition are re-classified into Domain 5.

Domain 5 (Includes Native Hawaiian or Pacific Islander): This domain includes any person marking the single race Native Hawaiian or Pacific Islander. It also includes the combination of Native Hawaiian or Pacific Islander and American Indian or Alaska Native not in IC. Also included is the combination of Native Hawaiian or Pacific Islander with Asian. All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are also included in Domain 5.

Domain 6 (Includes Non-Hispanic Asian): This domain includes any non-Hispanic person marking Asian as their single race. If a person self-identifies with Asian and American Indian or Alaska Native not in IC, they are included in Domain 6.

Domain 7 (Includes Non-Hispanic White or "Some other race"): Non-Hispanic White or Non-Hispanic "Some other race" persons are included Domain 7. Non-Hispanic persons who self-identify with American Indian or Alaska Native not in IC and are White or "Some other race" are classified into Domain 7. If a Native Hawaiian or Pacific Islander response is combined with a White or "Some other race" response, they also are included in Domain 7. A person who self-identifies with Asian and White or Asian and "Some other race" is also included in this domain. Finally, all non-Hispanic persons who self-identify with three or more races (excluding American Indian or Alaska Native in IC) are included in Domain 7. The only exception to this rule occurs when a Non-Hispanic White or Non-Hispanic "Some other race" person lives in Hawaii and classifies themselves as Native Hawaiian or Pacific Islander, regardless of whether they identify with other races. Persons who satisfy this criteria are re-classified into Domain 5.

Table 2: Census 2000 A.C.E. Post-stratification Domains for Hispanic

		Not in IC	Indian Country (IC)	
			Not On Res.	On Res.
Single race:				
American Indian or Alaska Native		3	2	1
Black		3	3	3
Native Hawaiian or Pacific Islander		3*	3	3
Asian		3	3	3
White		3	3	3
"Some other race"		3	3	3
American Indian or Alaska Native and:	Black	3	2	1
	Native Hawaiian or Pacific Islander	3*	2	1
	Asian	3	2	1
	White	3	2	1
	"Some other race"	3	2	1
Black and:	Native Hawaiian or Pacific Islander	3*	3	3
	Asian	3	3	3
	White	3	3	3
	"Some other race"	3	3	3
Native Hawaiian or Pacific Islander and:	Asian	3*	3	3
	White	3*	3	3
	"Some other race"	3*	3	3
Asian and:	White	3	3	3
	"Some other race"	3	3	3
American Indian or Alaska Native and:	Two or More Races	3*	2	1
All Else**		3*	3	3

* All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are included in Domain 5, which includes Native Hawaiian or Pacific Islander.

** All Else encompasses all remaining combinations which exclude American Indian or Alaska Native.

Table 3: Census 2000 A.C.E. Post-stratification Domains for Non-Hispanic

		Not in IC	Indian Country (IC)	
			Not On Res.	On Res.
Single race:				
American Indian or Alaska Native		2	2	1
Black		4	4	4
Native Hawaiian or Pacific Islander		5	5	5
Asian		6	6	6
White		7	7	7
"Some other race"		7	7	7
American Indian or Alaska Native and:	Black	4	2	1
	Native Hawaiian or Pacific Islander	5	2	1
	Asian	6	2	1
	White	7	2	1
	"Some other race"	7	2	1
Black and:	Native Hawaiian or Pacific Islander	4*	4	4
	Asian	4	4	4
	White	4	4	4
	"Some other race"	4	4	4
Native Hawaiian or Pacific Islander and:	Asian	5	5	5
	White	7*	7	7
	"Some other race"	7*	7	7
Asian and:	White	7	7	7
	"Some other race"	7	7	7
American Indian or Alaska Native and:	Two or More Races	7*	2	1
All Else**		7*	7	7

* All persons living in the state of Hawaii who classify themselves as Native Hawaiian or Pacific Islander, regardless of their Hispanic origin and whether they identify with a single or multiple race, are included in Domain 5, which includes Native Hawaiian or Pacific Islander.

** All Else encompasses all remaining combinations which exclude American Indian or Alaska Native.

ATTACHMENT: 1990 PES Post-Stratification

This attachment provides a brief summary of the 1990 PES post-stratification for the U.S. and Puerto Rico. Included below are the 51 post-stratum groups for the U.S. and the three post-stratum groups for Puerto Rico. Each of these post-stratum groups are further subdivided into the same seven age/sex groups.

Table 4a: 1990 PES 357 Design - 51 Post-Stratum Groups (U.S.)

Race/Hispanic Origin	Tenure	Urbanicity	N	M	S	W
Non-Hispanic White or Other	Owner	Large Urbanized Areas	1	2	3	4
		Other Urban	5	6	7	8
		Non-Urban	9	10	11	12
	Non-owner	Large Urbanized Areas	13	14	15	16
		Other Urban	17	18	19	20
		Non-Urban	21	22	23	24
Black	Owner	Large Urbanized Areas	25	26	27	28
		Other Urban	29			
		Non-Urban	30			
	Non-owner	Large Urbanized Areas	31	32	33	34
		Other Urban	35			
		Non-Urban	36			
Hispanic White or Other	Owner	Large Urbanized Areas	37	38	39	40
		Other Urban	41			
		Non-Urban	42			
	Non-owner	Large Urbanized Areas	43	44	45	46
		Other Urban	47			
		Non-Urban	48			
Asian or Pacific Islander	Owner		49			
	Non-owner		50			
Reservation Indians			51			

Table 4b: 1990 PES - 3 Post-Stratum Groups (Puerto Rico)

Place Type	
Central City in an MSA/PMSA	1
Non-central City in an MSA/PMSA	2
Not in an MSA/PMSA	3

Table 4c: 1990 PES - 7 Age/Sex Groups (U.S. & Puerto Rico)

	Male	Female
Under 18	A	
18 to 29	B	C
30 to 49	D	E
50+	F	G

Key:

MSA: Metropolitan Statistical Area, as defined by the Office of Management and Budget (OMB), will be referred to as MSA.

PMSA: Primary Metropolitan Statistical Area, as defined by the Office of Management and Budget (OMB), will be referred to as PMSA.



April 20, 2000

DSSD CENSUS 2000 PROCEDURES AND OPERATIONS MEMORANDUM Q-25

MEMORANDUM FOR: Documentation

Through: Donna L. Kostanich *DK*
Assistant Division Chief, Sampling and Estimation

From: Patrick J. Cantwell *PC*
Team Leader, Missing Data Team

Prepared By: Michael Ikeda
Missing Data Team

Subject: Accuracy and Coverage Evaluation Survey: Specifications for the Missing Data Procedures

A. Introduction

This document gives the procedures for handling missing data in the Census 2000 Accuracy and Coverage Evaluation (A.C.E.) sample. Section B provides general background. A noninterview adjustment procedure, in Section C, is used to account for whole-household nonresponse. A characteristic imputation procedure, in Section D, is used to assign values for specific missing demographic variables. Finally, persons with unresolved match, residence, or enumeration status have probabilities assigned based on the imputation cell estimation procedure in Section E.

The missing data procedures for the A.C.E. are similar to those used for the Integrated Coverage Measurement (ICM) sample in the Census 2000 Dress Rehearsal, although the cells used for imputation have been changed considerably. (See Section E.) An outline of the missing data procedures for the Dress Rehearsal ICM and a summary of related research is given in "Missing Data Procedures in the Census 2000 Dress Rehearsal Integrated Coverage Measurement Sample" by Michael Ikeda, Anne Kearney, and Rita Petroni (presented at the 1998 Meetings of the American Statistical Association). Additional information on variables used in the missing data processing can be found in the memorandum from P. Cantwell to M. Lynch "Census 2000: Data Requirements for A.C.E. Missing Data Input and Output Files." Additional information on variables and an overview of A.C.E. operations can be found in the memorandum from D. Childers to M. Ramos "Accuracy and Coverage Evaluation: The Design Document, DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01."

Direct any questions about this document to Michael Ikeda, SRD, 3211-4, x4864.

B. General Background

Census 2000 is conducted for the entire nation, Puerto Rico, and the outlying areas. For the Accuracy and Coverage Evaluation (A.C.E.), there are two separate samples: one for the 50 states and the District of Columbia, and a second sample for Puerto Rico. Although the National and Puerto Rico A.C.E. samples are processed separately by the A.C.E. missing data system, the A.C.E. missing data procedures are identical for the two samples. For simplicity, this document will describe the National sample.

The A.C.E. uses Dual System Estimation (DSE) to calculate estimates. The DSE procedure is to obtain a roster from the A.C.E. blocks independently of the Census. The independent roster (P Sample) and the Census roster (E Sample) are matched and the results of the matching are used to estimate the total number of persons in the population. Estimates are calculated separately for population subgroups called post-strata. Census 2000 uses a DSE method for handling movers, PES Procedure C (sometimes called simply PES C). Within each post-stratum PES Procedure C uses person in-movers to estimate the number of movers, and uses person out-movers to estimate the match rate among movers.

C. Noninterview Adjustment

Noninterview adjustment is only performed on the P-Sample. The noninterview adjustment procedure is almost identical to the procedures used in the Census 2000 Dress Rehearsal. There are two noninterview adjustments because of the use of PES Procedure C estimation. The two noninterview adjustments are identical to each other, except for the reference date for housing unit status. One noninterview adjustment is based on housing unit status as of Census Day. The other noninterview adjustment is based on housing unit status as of the day of A.C.E. interview.

Person nonmovers and person out-movers are used to determine Census Day housing unit status (that is, an interview or a noninterview). Person nonmovers and person in-movers are used to determine A.C.E. interview day housing unit status.

The noninterview adjustment based on Census Day status is used to adjust the weights of person nonmovers and person out-movers. The noninterview adjustment based on A.C.E. interview day status is used to adjust the weights of person in-movers.

Interview: A unit is an interview (for the given reference date) if there is at least one person (with name and at least two demographic characteristics) who possibly or definitely was a resident of the housing unit on the given reference date.

Noninterview: An occupied housing unit (as of the given reference date) that is not an interview is a noninterview.

The noninterview adjustment (for a given reference date) generally spreads the weights of

noninterviewed units equally over interviewed units in the same block cluster and similar type of basic address. Type of basic address has seven *possible* values in the Census 2000 A.C.E:

- 1 = one-family unit,
- 2 = multi-unit address,
- 3 = mobile home/trailer not in a mobile home park,
- 4 = mobile home/trailer in a park,
- 5 = one-family unit in a special place,
- 6 = multi-unit in a special place, and
- 7 = other.

For the noninterview adjustment, these seven values are collapsed as follows:

- 1 = one-family unit,
- 2 = multi-unit address, and
- 3 = all others.

If there are not enough interviewed units (as defined below) in the block cluster \times type of basic address category, then the weights of the noninterviewed units are spread out over a broader category of interviewed units.

The categories (in the order they are used) are as follows:

- 1) Block Cluster \times type of basic address category.
- 2) Re-coded A.C.E. Sample Stratum \times type of basic address category.
- 3) Block Cluster
- 4) Re-coded A.C.E. Stratum
- 5) State (DC and Puerto Rico are considered states for the noninterview adjustment)

Re-coded A.C.E. Sample Stratum, defined in this specification, is a classification of block clusters within each state for the purposes of the noninterview adjustment. The purpose is to group clusters with similar characteristics and sampling weights together. Two special sampling strata--one containing the original small block clusters and a second comprising the American Indian Reservations--are separate re-coded A.C.E. strata. For the remaining clusters, each demographic/tenure group cluster code (used during A.C.E. block cluster sampling) is a separate re-coded A.C.E. sample stratum. (See Attachment 4 for a description of the demographic/tenure code and the re-coded A.C.E. sample strata.)

For each category (except for State) there is a test to determine if there are enough interviewed units in the category. If the unweighted count of noninterviewed units in the given block cluster \times type of address category is more than twice the unweighted count of interviewed units in the current category of interviewed units, then we move to the next category of interviewed units. The weights of the noninterviewed units in the given block cluster \times type of basic address

category are spread out equally over units in the first category of interviewed units which contains enough interviewed units.

The actual noninterview adjusted weights (for a given reference date) are calculated as follows:

Let NI be the unweighted count of noninterviewed units in the given block cluster \times type of basic address category, W_{NI} be the sum of the initial weights of the noninterviewed units in the given block cluster \times type of basic address category, w_i be the initial weight for the i th unit, and A_j be the adjustment factor for the j th cluster \times type of basic address category. The initial weights are the P-Sample weights, reflecting the probability of selection at all stages of sampling (including within large blocks) and any potential trimming of the weights. A_j is initialized to 0 and modified as described below.

1) If NI is less than or equal to twice the unweighted count of interviewed units in the given block cluster \times type of basic address category, then add the following factor to A_j for the given block cluster \times type of place category:

$$\frac{W_{NI}}{\sum_{\text{interviews} \in B \times T} w_i}$$

where $B \times T$ indicates the given block cluster \times type of basic address category.

2) If NI is greater than twice the unweighted count of interviewed units in the given cluster \times type of basic address category but less than or equal to twice the unweighted count of interviewed units in the corresponding re-coded A.C.E. stratum \times type of basic address category then add the following factor to all A_j in the re-coded stratum \times type of basic address category:

$$\frac{W_{NI}}{\sum_{\text{interviews} \in R \times T} w_i}$$

where $R \times T$ indicates the re-coded A.C.E. stratum \times type of basic address category.

3) If NI is greater than twice the unweighted count of interviewed units in the given re-coded A.C.E stratum \times type of basic address category but less than or equal to twice the unweighted count of interviewed units in the given block cluster then add the following factor to all A_j in the block cluster:

$$\frac{W_{NI}}{\sum_{interviews \in B} w_i}$$

where B indicates the given block cluster.

4) If NI is greater than twice the unweighted count of interviewed units in the given block cluster but less than or equal to twice the unweighted count of interviewed units in the corresponding re-coded A.C.E. stratum then add the following factor to all A_j in the re-coded A.C.E. stratum:

$$\frac{W_{NI}}{\sum_{interviews \in R} w_i}$$

where R indicates the re-coded A.C.E. stratum.

5) If NI is greater than twice the unweighted count of interviewed units in the given re-coded A.C.E. stratum then add the following factor to all A_j in the corresponding state:

$$\frac{W_{NI}}{\sum_{interviews \in S} w_i}$$

where S indicates the state.

After the above steps have been done for every block cluster \times type of basic address category, the noninterview adjusted weight is calculated. For interviewed units in block cluster \times type of basic address category j , the noninterview adjusted weight is $w_i \cdot (1 + A_j)$. For noninterviewed units, the noninterview adjusted weight is 0. For vacant or deleted units, the noninterview adjusted weight is the same as the initial weight.

D. Characteristic Imputation

Characteristic imputation for the Accuracy and Coverage Evaluation P Sample is essentially identical to characteristic imputation for the Dress Rehearsal ICM. For the Census 2000 E Sample we use the demographic information from the Census 2000 Hundred-Percent Census edited file (HCEF). Therefore the only A.C.E. imputation that needs to be done in the E Sample is for E-Sample persons that cannot be matched to the HCEF. In the Dress Rehearsal all E-Sample persons matched to the HCEF (then called the CEF); we expect the same for the

Census 2000 E Sample. The methodology for any remaining E-Sample A.C.E. imputation is essentially the same as the P-Sample methodology.

The variables imputed in the A.C.E. are race, Hispanic origin, sex, tenure, and age. Each of these is necessary to place the A.C.E. sample person in the proper post-stratum. P-Sample person mover status is not considered when imputing characteristics. However, persons from the P-Sample whole-household out-mover interview path are considered to be a separate household for imputation purposes.

Distributions of age and sex for imputation are calculated nationally. Imputation for a specific missing characteristic is not affected by the imputation for other missing characteristics. For hot-deck imputation, the data are sorted using geographic and demographic variables, as well as household identifiers. (See Attachment 4 for the variables used and the sort order.) This essentially produces a geographic sort, except that block clusters with similar demographic characteristics will tend to be grouped together within a state.

Tenure: Tenure (collapsed to owner/non-owner) is imputed from the previous household with a similar type of basic address (structure code, in the E Sample) with tenure recorded. Type of basic address is collapsed in the same way as it was for the noninterview adjustment.

Race: Missing race is imputed from the distribution of reported race in the same household. That is, from the people in the same household who reported race, we randomly select one and assign the reported race value to the person with missing race.

If everyone in the same household is missing the value of race, then the distribution of the nearest previous household with reported race and the same value of re-coded Hispanic origin is used. The re-coded Hispanic origin variable takes one of three values: Non-Hispanic, Hispanic (that is, any of Mexican, Cuban, Puerto Rican, and other Hispanic origins), and missing origin. To see if the value of re-coded Hispanic origin is the same, we look at the first person on the roster of the nearest previous household--usually the householder.

If Hispanic Origin for the household (of the person with missing race) is missing, then the race distribution of the nearest previous household (regardless of Hispanic origin) with reported race is used. The flowchart in Attachment 1 illustrates the procedure.

Any of the 63 different combinations of the six basic race categories can be imputed. (The six categories are White, Black, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander, and Other.) All 63 categories are treated the same in the imputation; there are no special procedures for categories or groups of categories.

Hispanic Origin: Hispanic origin (collapsed to Non-Hispanic/Hispanic) is imputed from the distribution of reported Hispanic origin in the same household. If no one in the household has a nonmissing value of Hispanic Origin then the distribution of the nearest previous household with

reported Hispanic origin and similar race (race of the householder) is used. The race categories used in the imputation are: missing, white, other or white and other, remaining nonmissing categories. If race for the household is missing, then the Hispanic Origin distribution for the nearest previous household (regardless of race) with reported Hispanic Origin is used.

Age: We are actually imputing the age category, not the raw age. The age categories imputed are the same ones used in the post-stratification: 0-17, 18-29, 30-49, 50+. For most relationship to reference person categories in multi-person households, age category is imputed from the distribution of age category for persons with similar relationship to reference person, and same age category of reference person. For reference persons, the collapsed other relative category, and the collapsed nonrelatives category (or any category if age of reference person is missing), age category is imputed from the distribution of age category for persons with a similar relationship category in multi-person households. For the purposes of the age imputation, similar relationship is defined by the following collapsing of the original relationship to reference person categories: reference person, spouse, child, sibling, parent, other relative, nonrelative. If relationship is missing we impute from the distribution of age category in multi-person households. The flowchart in Attachment 2 illustrates the procedure for multi-person households. For one-person households, age category is imputed from the distribution of age category in one-person households.

Sex: Sex of reference person (with spouse present) or spouse of reference person is imputed by assigning the person with a missing value for sex the sex opposite to that of their spouse. If both reference person and spouse have sex missing, then sex for the reference person is imputed from the distribution of sex for reference persons with spouse present. The spouse is then assigned the sex opposite to that of the reference person. For one-person households, sex is imputed from the distribution of sex in one-person households. For the reference person (with no spouse present) of a multi-person household, the distribution of sex for reference persons of multi-person households with no spouse present is used. For persons (except reference persons and spouses) from multi-person households with non-missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons and spouses) from multi-person households. For persons from multi-person households with missing relationship, sex is imputed from the distribution of sex for persons (excluding reference persons) from multi-person households. The flowchart in Attachment 3 illustrates the procedure for multi-person households.

E. Assigning Match, Residence, and Correct Enumeration Probabilities

After A.C.E. follow-up is completed, for some people in the P Sample the final status for match or Census-Day residence or both is unresolved. Similarly, some people in the E Sample have an unresolved final enumeration status. For such people, a probability is assigned for the unresolved status by the imputation cell estimation (ICE) method. ICE calculates weighted ratios based on persons with resolved final status. For non-TES persons (that is, for whom the targeted extended search (TES) Person status = 0), we use the A.C.E. initial weights, reflecting

all stages of sampling and any potential weight trimming. For TES persons (TES Person status = 1) in TES clusters, we use the initial weights multiplied by the appropriate TES take-every. TES persons in clusters not in TES are given weights of 0, effectively dropping them from the calculations.

TES person status is assigned during person matching. P-Sample TES persons are persons (with sufficient information for matching) where the housing unit did not match during the initial housing-unit matching phase, and the household is a whole-household nonmatch when only within-cluster matching is considered. (For purposes of TES, housing units that did not go through initial housing-unit matching are considered nonmatches. This includes list/enumerate and re-listed clusters.) E-Sample TES persons are persons (with sufficient information for matching) where the housing unit was identified as a geocoding error in the housing unit phase.

ICE was also used in the Dress Rehearsal to estimate all three probabilities. We are, however, calculating the ratios at a more detailed level in Census 2000. We also have both certainty and non-certainty clusters in TES in Census 2000. In the Dress Rehearsal, we only had certainty TES clusters.

Attachment 5 gives more details on the before-followup and final match codes, and several other variables involved in the assignment of probabilities. Additional information can be found in the memorandum from P. Cantwell to M. Lynch "Census 2000: Data Requirements for A.C.E. Missing Data Input and Output Files" and the memorandum from D. Childers to M. Ramos "Accuracy and Coverage Evaluation: The Design Document, DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-01."

Residence Status. Persons with unresolved final Census-Day residence status are those persons with a final P-Sample match code of MU, NU, P, KI, or KP. (See attachment 5.) To assign probabilities of residence to these people, separate ratios are calculated within P-Sample match code *groups*. These groups are determined by the before-followup match code, initial whole/partial household match code, address code (HU match status from HU matching and conflicting household status), and person followup status. Seven P-Sample match code groups are defined:

- 1 = matches needing followup
- 2 = possible matches
- 3 = nonmatches needing followup from partial household nonmatches
- 4 = nonmatches needing followup from whole-household nonmatches (not conflicting households)
- 5 = nonmatches needing followup from conflicting households
- 6 = persons resolved before followup
- 7 = persons with insufficient information for matching

Persons resolved before followup are those persons with a before-followup match code of DP or

NC and those persons with a before-followup match code of M or NP who do not need followup.

Persons with insufficient information for matching (before followup) are those persons with a before-followup match code of KI or KP.

The residence probability for unresolved persons in groups 1-6 is the weighted proportion of persons in the given imputation cell who are residents. The residence probability for persons from group 7 is the weighted proportion of persons needing followup (in the same tenure and race group), that is, from groups 1-5 combined, who are residents. The weighted proportions are based on person non-movers and person out-movers with resolved final residence status, excluding persons with Computer Residence Status Code (A.C.E Status Code) of I (in-mover) or R (removed).

The imputation cells for estimation of P-Sample residence probability are defined below. Each internal table cell is an imputation cell.

Before Followup Group	Owner				Non-Owner			
	NH White		Others		NH White		Others	
1 = Matches needing followup (FU)								
2 = Possible matches								
3 = Partial HH nonmatches needing FU	V3a	V3b	V3a	V3b	V3a	V3b	V3a	V3b
4 = Whole HH nonmatches needing FU, not conflicting households								
5 = Nonmatches from conflicting HH								
6 = Resolved before FU								
7 = Insufficient information for matching	Weighted column average of BFU groups 1-5		Weighted column average of BFU groups 1-5		Weighted column average of BFU groups 1-5		Weighted column average of BFU groups 1-5	

Match code groups are defined above. Non-Hispanic White is defined by HISP2 = 1 (Non-Hispanic) and RACE = 32 (White, not in combination with other races). Tenure is defined by TENURE2: TENURE2 = 1 is Owner, TENURE2 = 2 is Non-Owner. V3 is a variable defined for group 3, partial household nonmatches. V3a includes persons in group 3 with AGE2 = 2 (18-29) and RELAT2 = 2 (Child of Reference Person). V3b includes all other persons in group 3.

The Census Day residence probability for person in-movers, identified using Computer Residence Status Code (A.C.E. status code), is irrelevant to estimation and is set to 0.

In the Dress Rehearsal, only three weighted ratios were calculated for residence probability: a ratio for persons sent to followup, a ratio for persons not needing followup, and an overall ratio

for persons with insufficient information for matching. Results from the Dress Rehearsal (Kearney and Ikeda, 1999) suggested that it would be useful to calculate separate ratios by match code group and to split persons from conflicting households into a separate match code group. The larger Accuracy and Coverage Evaluation sample size in Census 2000 allows us to separate matches needing followup from possible matches. Additional research and discussion suggested adding additional variables within match code group.

Match Status. For unresolved match status, the match probability for persons with unresolved match status is the proportion of matches in the same imputation cell among persons with resolved final match status (excluding confirmed Census Day nonresidents). Persons with unresolved match status are those persons with a final match code of P, KI, or KP. Most persons with unresolved match status are persons with insufficient information for matching. Note that all persons with unresolved match status also have unresolved residence status.

The imputation cells for estimation of P-Sample match probability are defined below. Each internal table cell is an imputation cell.

Mover Status	Housing-Unit Address Match Code			
	1 = HU match		2 or 4 (HU nonmatch or conflicting HH)	
Non-mover	0 imputes	1+ imputes	0 imputes	1+ imputes
Out-mover	0 imputes	1+imputes		

Mover status is determined by MOVERPER. MOVERPER = 1 indicates non-mover; MOVERPER = 3 indicates out-mover. The variable AMTIMP (0 imputes or 1+ imputes in the table) denotes how many of the following characteristics were imputed for the P-Sample person: age, sex, race, hispanic origin, and tenure. AMTIMP = 0 indicates no imputes.

The match probability is set to 0 for confirmed Census Day nonresidents. The match probability for person in-movers is irrelevant to estimation and is set to 0.

In the Dress Rehearsal, a single weighted ratio was calculated for match probability. Results from the Dress Rehearsal (Kearney and Ikeda, 1999) suggested that it would be useful to calculate separate ratios for person out-movers and person nonmovers. Additional research and discussion suggested adding additional variables within mover status.

Enumeration Status. Persons with unresolved enumeration status in the E Sample are those persons with final match codes of UE, MU, P, or GU. We assign them enumeration probabilities based on the weighted proportion of correct enumerations (among persons with resolved enumeration status) in the imputation cells as defined below. The cells are formed within E-Sample match code groups defined by before-followup match code, initial whole/partial household match code, address code (HU match status from HU matching and conflicting

household status), person followup status, and nonresponse followup universe status. Nine E-Sample match code groups are defined:

- 1 = matches needing followup
- 2 = possible matches
- 3 = nonmatches from partial household nonmatches
- 4 = nonmatches from whole-household nonmatches where the HU matched in initial HU matching (not conflicting households)
- 5 = nonmatches from conflicting households where the E-Sample HU was not in regular nonresponse followup (NRFU)
- 6 = nonmatches from conflicting households where the E-Sample HU was in regular NRFU (NRU=3)
- 7 = nonmatches from whole-household nonmatches where the HU did not match during initial HU matching
- 8 = persons resolved before followup
- 9 = persons with insufficient information for matching.

For the purposes of match code group assignment, nonmatches are those persons with a before-followup match code of NE, GS, GC, or GU. Persons resolved before followup are those persons with a before-followup match code of DE, GE, or FE and those persons with a before-followup match code of M who do not need followup. Persons with insufficient information for matching (before followup) are those persons with a before-followup match code of KE. Note that E-Sample persons with insufficient information for matching are treated as erroneous enumerations.

The imputation cells for estimation of E-Sample correct enumeration probability are defined below. Each internal table cell is an imputation cell.

Before Followup Group	0 Imputes		1+ Imputes	
1 = Matches needing followup (FU)				
2 = Possible matches				
3 = Partial household (HH) nonmatches	V3a	V3b	V3a	V3b
4 = Whole HH nonmatches where HU matched; not conflicting households	NH White	Others		
5 = Nonmatches from conflicting HH; HH not in regular NRFU				
6 = Nonmatches from conflicting HH; HU in regular NRFU				
7 = Whole HH nonmatches, HU did not match in HU matching	NH White	Others		
8 = Resolved before FU	NH White	Others		
9 = Insufficient information for matching				

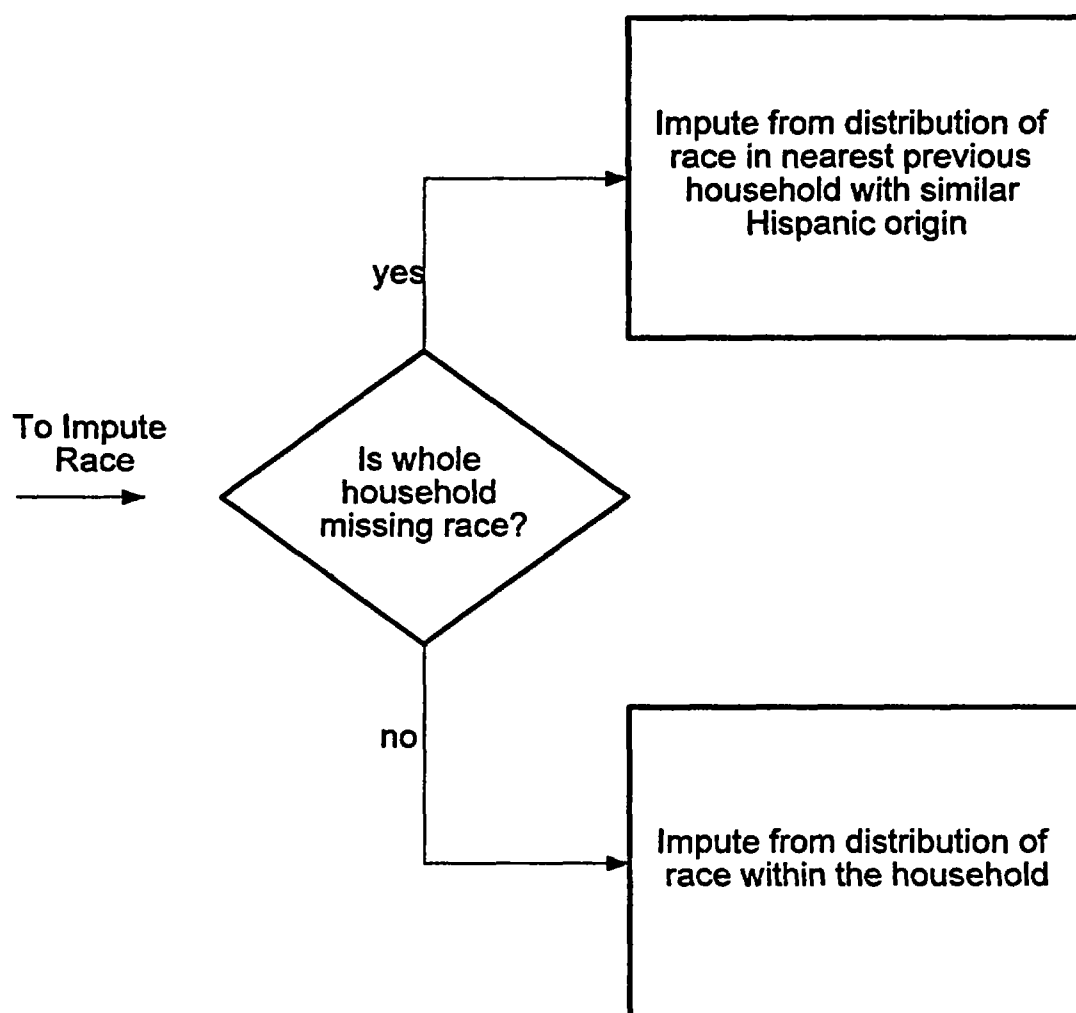
Match code groups are defined just above. Non-Hispanic White is defined by PSPAN = 1 (Non-Hispanic) and RACE = 32 (White). 0 imputes is defined by AMTIMP = 0 (no variables imputed in either E-Sample or HCEF). AMTIMP considers age, sex, race, hispanic origin, and tenure. Age is not considered to be imputed if either age or date of birth was collected (and do not contradict each other). V3 is a variable defined for match code group 3 (partial household nonmatches). V3a includes persons in match code group 3 with AGE2 = 2 (18-29) and RELAT2 = 2 (Child of Reference Person). V3b includes all other persons in match code group 3.

In the Dress Rehearsal, matches needing followup were grouped with possible matches. In addition, E-Sample match code groups 4-6 were a single group in the Dress Rehearsal. Results from the Dress Rehearsal (Kearney and Ikeda, 1999) suggested that it would be useful to separate conflicting households from other whole-household nonmatches where the address matches and to further separate NRFU conflicting households from non-NRFU conflicting households. The larger Accuracy and Coverage Evaluation sample size in Census 2000 allows us to separate matches needing followup from possible matches. Additional research and discussion suggested adding additional variables within match code group.

Special Cases

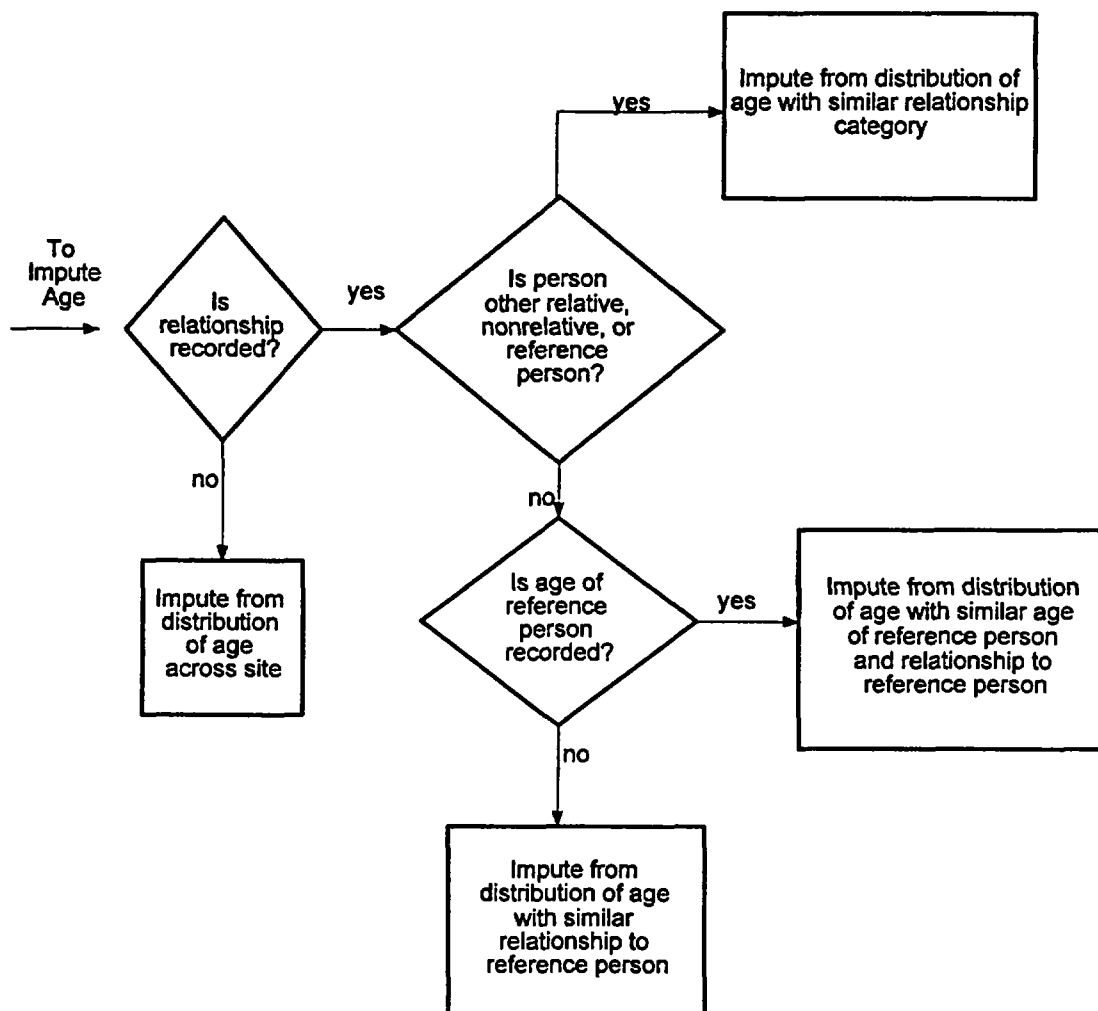
1. There is an additional adjustment made due to (i) duplication with persons subsampled out of the E-Sample (in large clusters) and (ii) duplication with Census group quarters persons in the same Accuracy and Coverage Evaluation block cluster. If an E-Sample person is duplicated with K persons who are either subsampled out of the E-Sample or from Census group quarters, then the initial correct enumeration probability is multiplied by $1/(K+1)$, since we do not know which person is the "real" person.
2. For assigning probabilities through ICE for unresolved status of each type, we have selected the variables to define imputation cells in such a way that we expect at least several hundred unweighted resolved cases in each cell. Regardless of the resulting frequencies, we will *not* collapse any cells. If there are no persons in a cell--no one resolved and no one unresolved--there is no need to assign probabilities in that cell. If there are no resolved persons in a cell *and* at least one unresolved person in the cell, we will assign the unresolved persons in that cell a probability of 1.0. We do not expect the latter situation to occur.

Imputation of Race for the P Sample

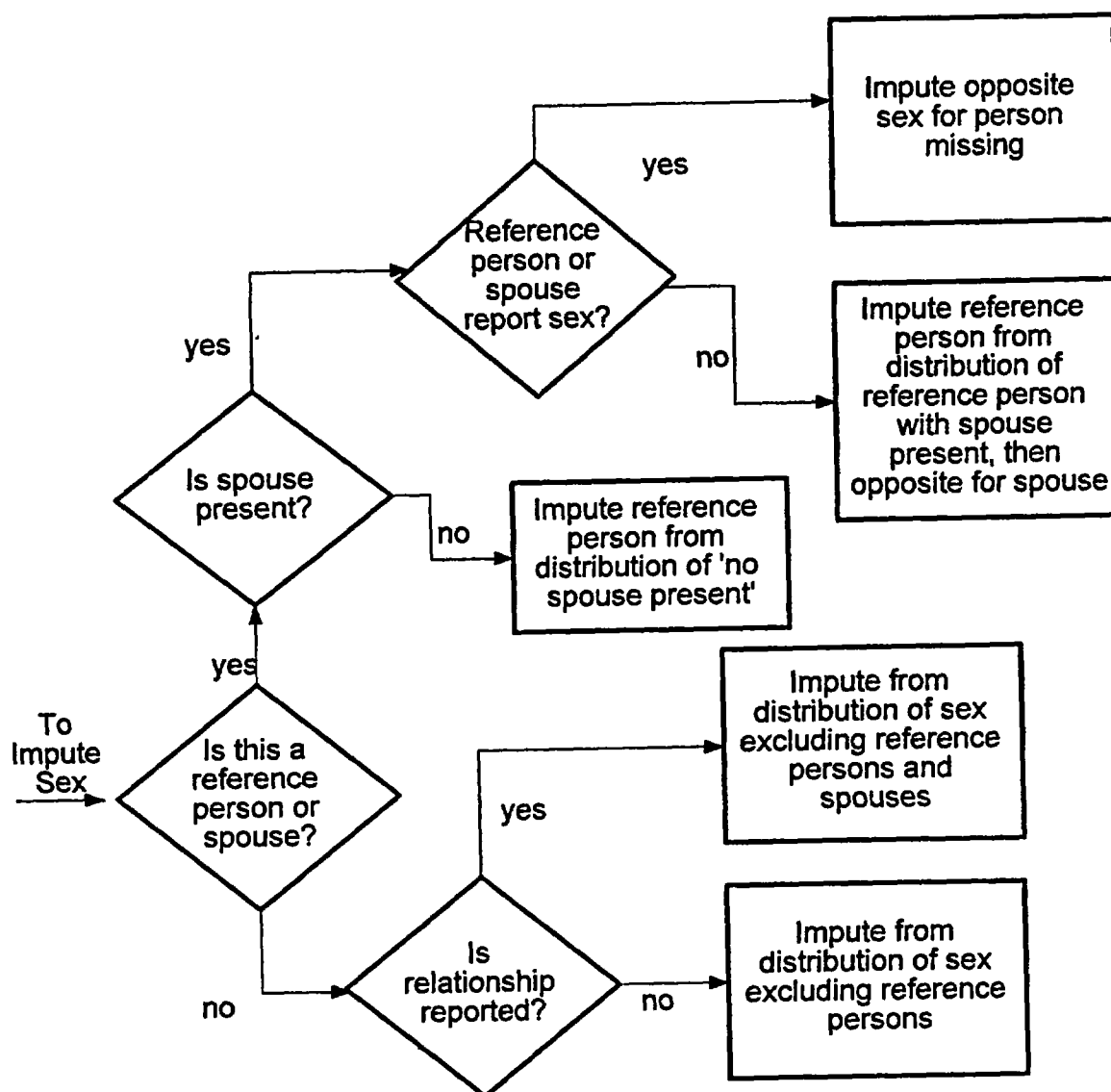


Note that if whole household is also missing Hispanic origin, then we impute from the distribution of race in the nearest previous household with reported race (regardless of Hispanic origin).

Imputation of Age Category for the P Sample (Multi-Person Households)



Imputation of Sex for the P Sample (Multi-Person Households)



Demographic/Tenure Code

The Demographic/Tenure Code is a block cluster level variable used in block cluster sampling. The demographic/tenure codes are based on the estimated demographic makeup of the cluster (based on 1990 Census data) and are as follows:

- 1 = Pacific Islander Renter
- 2 = Pacific Islander Owner
- 3 = Asian Renter
- 4 = Asian Owner
- 5 = American Indian/Alaska Native Renter
- 6 = American Indian/Alaska Native Owner
- 7 = Hispanic Renter
- 8 = Hispanic Owner
- 9 = Black Renter
- 10 = Black Owner
- 11 = All Other Renter
- 12 = All Other Owner

Codes 1-12 are not used for Puerto Rico. Instead the following codes are used:

- 13 = Renter
- 14 = Owner

A more detailed description of the Demographic/Tenure code can be found in the memorandum from D. Kostanich to D. Stoudt, "Accuracy and Coverage Evaluation (ACE) Survey: Universe File and Block Cluster Sampling Parameter File Specification, DSSD Census 2000 Procedures and Operations Memorandum Series R-5."

Re-Coded A.C.E. Sample Stratum Codes

- 1 - 14: = demographic/tenure code (above), when sampstr = 2 or 3 (these are the sampling stratum codes denoting medium and large blocks, respectively)
- 15: if sampstr = 1 (this is the sampling stratum code denoting small block clusters)
- 16: if sampstr = 4 (this is the sampling stratum code denoting American Indian Reservations)

Variables Used in the Sort for Characteristic Imputation

Within variables, the sort is ascending on the numerical values.

Sorting the P Sample for Imputation

Region

State

American Indian Country indicator

Demographic/tenure group code

Block cluster

Household identifiers:

map spot number

within-map spot unit ID

MOVERHH (a household mover flag, indicating whether the person was from the
out-mover path)

RELATE (relationship to reference person).

Sorting the E Sample for Imputation

Region

State

American Indian Country indicator

Demographic/tenure group code

Block cluster

Household identifiers:

census unit ID

PRELSUP (relationship to reference person).

SELECTED VARIABLES USED IN ASSIGNMENT OF PROBABILITIES

Match Codes from Person Matching

Before Followup Match Codes

M	=	The P-sample and census people match.
P	=	The P-sample and census people are possible matches. Additional follow-up is needed.
NP	=	The P-sample person is not matched to the census.
NE	=	The E-sample person is not matched to a P-sample person, additional follow-up is needed.
KP	=	Match not attempted for the P-sample person, because the name is blank or incomplete or not a valid name.
KE	=	Match not attempted for the E-sample person. The E-sample name is blank or incomplete or not a valid name.
DP	=	The P-sample person is a duplicate of another P-sample person.
DE	=	The E-sample person is a duplicate of another E-sample person or a duplicate of a census person in a surrounding block (DE is also assigned to non E-Sample persons duplicated with E-Sample persons in the same cluster).
GE	=	The E-Sample person is erroneously enumerated in this block cluster, because the housing unit is a geocoding error.
FE	=	The E-Sample person is a dog, cat, or other animal that should not go to follow-up. The enumeration is fictitious.
NC	=	The P-sample nonmatch was found on the census roster. This person in a partial nonmatch household was not matched to the census because only name was collected in the census for this person in a large household and the census person was not data defined. No follow-up interview is necessary.
GS	=	The E-sample person is enumerated in a housing unit that exists in a surrounding block.
GC	=	The E-Sample person is enumerated in a housing unit that exists in the sample cluster.
GU	=	The geographic work for the targeted extended search is unresolved. It is not clear where the housing unit is located.

Final P-Sample Match Codes

Matched—Confirmed Resident

M	=	The P-sample and the census people were matched.
MR	=	The P-sample follow-up interview determined that the matched person with unresolved residence status is a resident as of census day.

Matched—Unresolved Residence Status

MU	=	The A.C.E. person follow-up interview obtained no useful information to resolve the residence status for the matched person who had a residence status of unresolved before follow-up. The P-sample person's residence status is unresolved.
----	---	--

Not Matched—Confirmed Resident

NP	=	The P-sample person is not matched to an E-sample person. There was no follow-up for the whole household nonmatches from person interviews with household members and the whole household nonmatches were not conflicting household nonmatches. The P-Sample person is considered to be a resident on census day.
----	---	---

- NC = The P-sample nonmatch was found on the census roster. This person in a partial nonmatch household was not matched to the census because only name was collected in the census for this person in a large household and the census person was not data defined. No follow-up interview is necessary. The P-Sample person is considered to be a resident on census day.
- NR = The P-sample person is identified as a resident in the block cluster on census day during the A.C.E. person follow-up interview.

Not Matched—Unresolved Residence Status

- NU = Not enough information is collected during the A.C.E. person follow-up interview to identify the P-sample person as a resident or nonresident in the block cluster. The match status for the P-sample person is nonmatch.

Unresolved Match and Residence Status

- P = There is not enough information collected to determine if the possible match is a match or not. The match and residence status of the P-sample person are unresolved.
- KI = Match not attempted for the P-sample person because the person has insufficient information for matching and follow-up. The name is blank or incomplete or the name is complete but the person characteristic data are not sufficient (one or no person characteristics provided). This is a computer assigned code only. Both the match and residence status of the P-Sample person are unresolved.
- KP = Match not attempted for the P-sample person. The name is blank or incomplete or not a valid name. Both the match and residence status of the P-Sample person are unresolved.

Removed from the P-sample

- FP = The P-sample person is fictitious in this block cluster. The person is included in the independent roster in error during the CAPI interview. This person is not included in the final list of P-sample people.
- NL = The P-Sample person did not live at the sample address or in the block cluster on census day and was listed as a nonmover or out-mover in error. This person is removed from the list of P-Sample people since the person was collected during the person interview in error.
- NN = The P-sample person is identified as a nonresident in the block cluster on census day during the A.C.E. person follow-up interview, because the person lived in group quarters or had another residence where the person should have been counted on census day according to census residence rules. This person is removed from the list of P-sample people, since he or she was collected during the person interview in error.
- GP = The P-Sample person is removed because the person interview was conducted at a housing unit that exists outside the sample block cluster. The person follow-up identified this housing unit as a P-Sample geocoding error.
- DP = The P-sample person is a duplicate of another P-sample person.
- MN = The A.C.E. person follow-up interview determined that the matched person with unresolved residence status is not a resident in this housing unit or in this block cluster. The person is no longer in the list of P-sample people.

Final E-Sample Match Codes

Correctly Enumerated

- M = The P-sample and E-sample people were matched. The E-sample person is correctly enumerated.
- CE = The E-sample nonmatch is identified as correctly enumerated during the A.C.E. person follow-up interview.
- MR = The A.C.E. person follow-up interview determined that the matched person with unresolved residence status is a resident. The E-Sample person is a correct enumeration.

Erroneously Enumerated

- GE = The E-sample person is erroneously enumerated in this block cluster, because the census housing unit is a geocoding error (i.e., counted in the block cluster in error). The E-sample person should have been enumerated elsewhere in the census.
- EE = The E-sample nonmatch is identified as erroneously enumerated from the A.C.E. person follow-up interview.
- FE = The E-sample nonmatch is determined to be fictitious in this block cluster during the follow-up interview. The person may have existed, but should not have been enumerated in the census within this block cluster. The E-sample person is erroneously enumerated in the census in this block cluster.
- DE = The E-sample person is a duplicate of another E-sample person or a duplicate of a census person in a surrounding block (DE is also assigned to non E-Sample persons duplicated with an E-Sample person in the same cluster).
- MN = The A.C.E. person follow-up interview determined that the matched person with unresolved residence status is not a resident in this housing unit or in this block cluster. The E-sample person is an erroneous enumeration.
- KE = Match not attempted for the E-sample person. The name is blank or incomplete or the name is complete but there are one or no person characteristics (computer assigned). The census name is blank or incomplete or not a valid name (clerically assigned).

Unresolved

- UE = Not enough information is collected during the A.C.E. person follow-up interview to identify the E-sample person as correctly or erroneously enumerated in the E-sample. The enumeration status for the E-sample person is unresolved.
- MU = The A.C.E. person follow-up interview obtained no useful information to resolve the residence status for the matched person with unresolved residence status. The E-sample person's enumeration status is unresolved.
- P = There is not enough information collected to determine if the possible match is a match or not. The enumeration status of the E-sample person is unresolved.
- GU = The geographic work for the targeted extended search is unresolved. It is not clear where the housing unit is located.

Other Variables

Computer Residence Status Code (A.C.E. Status Code) (P-Sample)

N = Nonmover

O = Out-mover

U = Unresolved Residence Status (Can be either Nonmover or Out-mover)

I = In-mover

R = Remove (Census Day nonresident)

Persons with status codes of I or R have blank person match codes and are automatically assigned residence and match probabilities of 0.

Person Mover Flag (P-Sample)

1 = Nonmover

2 = In-mover

3 = Out-mover

DSE Follow-up Flag (P-Sample and E-Sample)

blank = No one in household needs follow-up

0 = Someone in household needs follow-up but person doesn't

1,2 = Person needs followup

Initial Whole/Partial Match Code (P-Sample and E-Sample)

1 = Partial Household Match

2 = Whole household nonmatch

3 = Whole household match

4 = Other

Persons with before-followup match codes of P, KI, KP, and DP are ignored when determining the initial whole/partial match code.

P-Sample Address Code

1 = HU Matched during HU Phase

2 = HU not matched during HU Phase

4 = Conflicting Households

E-Sample Address Code

1 = HU Matched during HU Phase

2 = HU not matched during HU Phase

3 = HU added to DMAF after HU Phase of A.C.E.

4 = Conflicting Households

Address code values of 2-3 are considered to be "HU not matched during HU matching" for the purposes of match code group assignment.

NRU (Non-Response Follow-Up Universe) (E-Sample)

- 0 = universe not set
- 1 = Not in NRFU, data received
- 2 = Not in NRFU but NRD, NRS, NRC, and NRPOP were set on DMAF by Update/Enumerate or List/Enumerate
- 3 = in NRFU, Nonresponse
- 4 = in NRFU, too late for mailout

TES Person Status (P-Sample and E-Sample)

- 0 = Not TES Person
- 1 = TES Person

TENURE2 (Re-coded Tenure) (P-Sample and E-Sample)

- 1 = Owner
- 2 = Renter

AGE2 (Age Category) (P-Sample and E-Sample)

- 1 = 0-17
- 2 = 18-29
- 3 = 30-49
- 4 = 50+

P-Sample AMTIMP

- 0 = No Imputation
- 1 = At least one variable (age, race, sex, hispanic origin, or tenure) imputed

E-Sample AMTIMP

- 0 = No Imputation
- 1 = At least one variable (age, race, sex, hispanic origin, or tenure) imputed in E-Sample
- 2 = At least one variable imputed on HCEF (age is not considered to be imputed if either age or date of birth was collected, other variables are imputed if they are edited or allocated)

RELAT2 (Re-coded Relationship) (P-Sample and E-Sample)

- 0 = Missing
- 1 = Spouse
- 2 = Child
- 3 = Sibling
- 4 = Parent
- 5 = Other Relative
- 6 = Nonrelative
- 9 = Reference Person

Hispanic Origin (HISP2 in P-Sample, PSPAN in E-Sample)

- 1 = Non-Hispanic
- 2 = Hispanic

Race (P-Sample and E-Sample)

1=Other Race
2=Pacific Islander (PI)(includes Native Hawaiians)
3=PI and Other Race
4=Asian
5=Asian and Other Race
6=Asian and PI
7=Asian, PI and Other Race
8=American Indian or Alaska Native (AI)
9=AI and Other Race
10=AI and PI
11=AI, PI, and Other Race
12=AI and Asian
13=AI, Asian, and Other Race
14=AI, Asian, and PI
15=AI, Asian, PI, and Other Race
16=Black
17=Black and Other Race
18=Black and PI
19=Black, PI, and Other Race
20=Black and Asian
21=Black, Asian, and Other Race
22=Black, Asian, and PI
23=Black, Asian, PI, and Other Race
24=Black and AI
25=Black, AI, and Other Race
26=Black, AI, and PI
27=Black, AI, PI, and Other Race
28=Black, AI, and Asian
29=Black, AI, Asian, and Other Race
30=Black, AI, Asian, and PI
31=Black, AI, Asian, PI, and Other Race
32=White
33=White and Other Race
34=White and PI
35=White, PI, and Other Race
36=White and Asian
37=White, Asian, and Other Race
38=White, Asian, and PI
39=White, Asian, PI, and Other Race
40=White and AI
41=White, AI, and Other Race
42=White, AI, and PI
43=White, AI, PI, and Other Race
44=White, AI, and Asian
45=White, AI, Asian, and Other Race
46=White, AI, Asian, and PI
47=White, AI, Asian, PI, and Other Race
48=White and Black
49=White, Black, and Other Race
50=White, Black, and PI
51=White, Black, PI, and Other Race
52=White, Black and Asian
53=White, Black, Asian, and Other Race
54=White, Black, Asian, and PI
55=White, Black, Asian, PI, and Other Race
56=White, Black, and AI
57=White, Black, AI, and Other Race
58=White, Black, AI, and PI
59=White, Black, AI, PI, and Other Race
60=White, Black, AI, and Asian
61=White, Black, AI, Asian, and Other Race
62=White, Black, AI, Asian, and PI
63=White, Black, AI, Asian, PI, and Other Race